UNIVERSITY OF CALIFORNIA
RIVERSIDE

Automated Analysis of User-Generated Content on the Web

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

by

Ryan Rivas

March 2021

Dissertation Committee:
Dr. Vagelis Hristidis, Chairperson
Dr. Eamonn Keogh
Dr. Amr Magdy
Dr. Vagelis Papalexakis

The Dissertation of Ryan Rivas is approved:

Committee Chairperson

University of California, Riverside

The text of this dissertation, in part, is a reprint of the material as it appears in the following publications:

Shetty P, Rivas R, Hristidis V. Correlating ratings of health insurance plans to their providers' attributes. Journal of Medical Internet Research. 2016;18(10):e279.

Rivas R, Montazeri N, Le NX, Hristidis V. Automatic classification of online doctor reviews: evaluation of text classifier algorithms. Journal of Medical Internet Research. 2018;20(11):e11141.

Rivas R, Patil D, Hristidis V, Barr JR, Srinivasan N. The impact of colleges and hospitals to local real estate markets. Journal of Big Data. 2019 Dec 1;6(1):7.

Rivas R, Sadah SA, Guo Y, Hristidis V. Classification of health-related social media posts: evaluation of post content classifier models and analysis of user demographics. JMIR Public Health and Surveillance. 2020;6(2):e14952.

The coauthor Vagelis Hristidis listed in these publications directed and supervised the research which forms the basis for this dissertation.

ABSTRACT OF THE DISSERTATION

Automated Analysis of User-Generated Content on the Web

by

Ryan Rivas

Doctor of Philosophy, Graduate Program in Computer Science
University of California, Riverside, March 2021
Dr. Vagelis Hristidis, Chairperson

Social media users generate large volumes of data every day. Analysis of this data is an important tool in several areas. For example, the study of users' opinions, behaviors, and topics of discussion can be of use in the field of health care. The first part of my research involves using existing tools to perform analysis of Web content. Specifically, it first compares between health care provider attributes and quality measures of the insurance plans they accept. This is followed by analyses of how real estate prices and related metrics are affected by proximity to a university or hospital. Further, this research studies user behaviors and discussion topics to find differences in how various demographic groups generate content on health-related Web forums and on health-related discussions in general social media. The remainder of this dissertation shifts its focus from analysis of Web content to proposing new tools to perform similar analyses. It first proposes and evaluates natural language processing-based methods to automatically classify patient opinions in doctor reviews. This work also introduces a variant of the review classification problem where class labels can represent two opposing opinions that are not necessarily positive or negative. This is followed by an exploration of methods to effectively filter social media posts according to a user’s interests. The key challenge behind this work is to determine how to use this information to maximize a trained text classification model’s performance in classifying new posts. Finally, this dissertation proposes a multimodal Twitter embedding model that can leverage information from several parts of a tweet, such as text, image, and location. Such a model can have several applications for both researchers and Twitter users without the need to train a separate model for each application.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Web 2.0 refers to the World Wide Web as a source of user-generated content. This differs from the early years of the Web, in which content creators were few and users were primarily consumers of Web content. Web 2.0 has existed for several years now, and many websites primarily focus on content generated by their users. This includes general social media sites such as Facebook and Twitter, content aggregators like Reddit, and review sites Yelp and Healthgrades. Even websites that do not primarily focus on user-generated content, including Amazon and many news websites, also include user-generated content in the form of product reviews and comments, respectively. Users of these websites generate huge amounts of data every day, and these datasets present valuable targets for analysis and informative training data for machine learning models. Considering businesses and researchers alike can gain useful insights from this data, it is clear that analysis of user-generated content is a valuable tool in the Web 2.0 landscape. New analyses, as well as new methods to conduct those analyses, are always sought by those who might reap their benefits. To that end, this dissertation presents analyses of Web content as well as new methods that can be applied to new analyses to further learn from the content generated by Internet users.

1.2 Research Problems

This dissertation is composed of two sections that represent two different approaches to the analysis of Web content. The first section consists of Chapters 2, 3, and 4 and presents primarily statistical analyses of health insurance plan ratings, real estate prices, and health-related discussions. These chapters present results based on data collected from static periods of time, but their methodologies can be automated to gain new insights as new data is collected and analyzed over time. The second section, which contains Chapters 5, 6, and 7, presents new problems in applying machine learning to user-generated Web content and proposes and evaluates methods to address them. The remainder of this chapter summarizes the research presented in the remaining chapters of this dissertation.

Correlating Ratings of Health Insurance Plans to Their Providers' Attributes

This study, motivated by the push for quality measures in health care, examined the relationship between health insurance plan quality measures and the attributes of health care providers in an insurance plan's network. Insurance plan quality measures collected from NCQA included measures for overall quality (e.g. rank), customer service (e.g. satisfaction with physicians), prevention (e.g. cancer screening), and treatment (e.g. diabetes treatment). Doctor attributes, which included patient ratings, hospital affiliations, number of referrals, and several other measures, came from several sources: doctor review websites Vitals and Healthgrades, the Centers for Medicare and Medicaid Services, and U.S. News & World Report.

A key challenge in this study was aggregating the data from these sources. Insurance plan names, hospital specialties, and doctor information may differ between the various sources in which they appear. Mappings between sources solved this problem. Generating these mappings required different approaches for each type for data, e.g. to match doctors across sources a custom-made algorithm generated an overall similarity score based on the similarity of individual doctor attributes.

The findings of this study may provide new insights to patients, insurers, and health care providers. The analysis of the relationship between health insurance plans and health care provides attributes is presented in Chapter 2.

The Impact of Colleges and Hospitals to Local Real Estate Markets

This study examined how home price and rent are affected by the presence of a university or hospital nearby. We collected data from several sources, including Zillow, U.S. News & World Report, and the Centers for Medicare and Medicaid Services to analyze home price and rent at both the ZIP code level (i.e. the median home price and rent in each ZIP code containing a university or a hospital) and the level of individual homes.

At the ZIP code level, we looked at home price, rent, appreciation, volatility, and vacancies over time in ZIP codes with a university or a hospital. We compared groups of ZIP codes separated by several measures, such as university/hospital size, population, and population density. At the level of individual homes, we looked for correlations between home price/rent and distance from a university or hospital based on house size, university rank, and other factors.

The results of this study generally agreed with our expectations, but we found some surprising results as well. The analysis of how universities and hospitals affect real estate prices is presented in Chapter 3.

Classification of Health-Related Social Media Posts: Evaluation of Post Content Classifier Models and Analysis of User Demographics

This study was motivated by the rising volume of health-related social media activity. Its objective was to classify the content of health-related social media posts and observe the differences in post content among several user demographics. We analyzed posts from both health-related Web forums (WebMD and DailyStrength) and general social media (Twitter and Google+).

Our analysis was based on identifying post content categories, e.g. sharing experiences and asking for medical advice. We trained and evaluated text classifiers for this task and used the best-performing models to classify additional posts for further analysis. We then compared the frequencies of these post content categories to the demographics of the authors of posts containing these categories.

The results of this study provide useful information that can help health care providers, researchers, and health advocates reach the right demographic groups. The process and results of analyzing the demographics of health-related social media users in terms of the content of their posts are further detailed in Chapter 4.

Automatic Classification of Online Doctor Reviews: Evaluation of Text Classifier Algorithms

In this study, we present the doctor review classification problem. Given a dataset of doctor review sentences and a class c, each sentence may be classified as one of three possible values: neutral (i.e. the sentence is unrelated to c), or one of two opposing opinions for c, e.g. the doctor spent either a long time or a short time with the patient. These opposing opinions are not strictly positive or negative, as different people may have different views on whether "long time" and "short time" correspond to "good" and "bad." To explore this problem, we created a dataset of doctor review sentences from Vitals and labeled them with several opinion classes.

Our experiments evaluated three distinct types of machine learning algorithms. These included traditional methods such as random forests, deep learning methods such as convolutional neural network, as well as methods based on natural language processing methods. These included both previous work and a proposed method that generates a list of rules based on syntactic dependency tree patterns extracted from training data.

The performance of all methods evaluated was low on average, but the results show the feasibility of addressing the doctor review classification problem. An improved model could be used to allow patients to search for doctors based on their personal preferences. The details of the doctor review classification problem and the methods evaluated to address it are presented in Chapter 5.

Effective Social Post Classifiers on Top of Search Interfaces

This study presents the problem of retrieving training data for a social media post classifier from behind a constrained search interface. The motivation behind addressing this problem is creating a "personal classifier" to filter social media posts according to a user's content preferences. Such a system would query the user for a few keywords relevant to a topic of interest, return posts for labeling (which may be implicit, e.g. via clickthroughs), then train a classifier and use it to filter future posts. The goal in addressing this problem is thus to create a method that retrieves a dataset capable of training a classifier with high accuracy.

However, there are some challenges in addressing this problem. Keywords provided by the user are not perfect, i.e. they do not always retrieve relevant results. To train a classifier with high accuracy, the training dataset generated should be both balanced (as close to 50% positives and 50% negatives as possible) and diverse (covers a wide range of both positives and negatives). Finally, the aforementioned constrained search interface limits how many posts we can retrieve in a given amount of time and prevents the use of active learning since we cannot access all of the data.

To address this problem and the challenges thereof, we created several methods to retrieve a set of social media posts for training a classifier. Our best method obtains samples of posts for the input keywords and uses their labels to determine which keywords to use to retrieve additional posts. It also retrieves random posts to prevent "noise" generated by keyword-based negatives in the dataset and to achieve better balance.

Our experiments use data from Reddit, DailyStrength, and The Huffington Post to show that our proposed method outperforms all other methods evaluated. Our results also show the importance of balance and diversity in creating a dataset for training a classifier. The training post retrieval problem over constrained search interfaces, as well as the proposed method to address it, are further explained in Chapter 6.

Holistic Embedding Generation for Twitter Machine Learning Applications

The purpose of this study was to determine whether a joint embedding model can take advantage of multiple components of a tweet to both achieve better performance on existing tasks normally involving joint embeddings and perform well in new tasks. Current joint embeddings typically only incorporate two to three modalities, e.g. text and image for image-text retrieval. However, a model that also uses additional components of a tweet, such as author and hashtags, has the potential to leverage this information to achieve better results. Additional components in the joint embedding space may also open up opportunities to apply the model to other tasks that typically do not call for joint embeddings such as hashtag recommendation and location prediction.

Our proposed method extends previous work on multimodal joint embeddings for image-text retrieval to incorporate five tweet components and put them in a joint embedding space such that the distance between components from the same tweet should be minimized. Besides images and text, the components of a tweet we consider also include hashtags, the author of the tweet and the location of a post (represented by the text of tweets from that location). We learned a joint embedding model through a deep neural framework that uses word embeddings, graph embeddings, a convolutional neural network for images, and a recurrent neural network for text to learn a joint embedding from the tweet components.

We tested the model on several tasks, including image-text retrieval, hashtag recommendation, bot detection, and location prediction, and compared our results to baselines specific to each of these tasks. We present these results, as well as the details of our proposed framework, in Chapter 7.

Chapter 2

Correlating Ratings of Health Insurance Plans to Their Providers' Attributes

Background: There is a push towards quality measures in health care. As a consequence, the National Committee for Quality Assurance (NCQA) has been publishing insurance plan quality measures.

Objective: The objective of this study was to examine the relationship between insurance plan quality measures and the participating providers (doctors).

Methods: We collected and analyzed provider and insurance plan data from several online sources, including provider directories, provider referrals and awards, patient reviewing sites, and hospital rankings. The relationships between the provider attributes and the insurance plan quality measures were examined.

Results: Our analysis yielded several findings: (1) there is a moderate Pearson correlation ($r = 0.376$) between consumer satisfaction insurance plan scores and review ratings of the member providers, (2) referral frequency and provider awards are negligibly correlated to consumer satisfaction plan scores (correlations of $r = 0.031$ and $r = 0.183$, respectively), (3) there is weak positive correlation ($r = 0.266$) between the cost charged for the same procedures and consumer satisfaction plan scores, and (4) there is no significant correlation between member specialists' review ratings and specialty-specific insurance plan treatment scores for most specialties, except a surprising weak negative correlation for diabetes treatment ($r = -0.259$).

Conclusions: Our findings may be used by consumers to make informed choices about their insurance plans or by insurances to understand the relationship between patients' satisfaction and their network of providers.

2.1 Introduction

There are several health insurance marketplaces and search portals (e.g. ehealthinsurance.com) that help individuals and small employers shop for, select, and enroll in high-quality, affordable health plans. Insurance plans are generally ranked based on relative quality and price. These marketplaces and search portals need to establish criteria and selection processes for quality measures. Most of them measure the quality of health plans by surveying plan enrollees on their satisfaction with their coverage and then publishing quality and satisfaction data online [1]. However, the relationship between the quality of insurance plans and the properties of providers in their networks has not been adequately studied, which is the focus of this study.

We collected a rich set of data for each provider ranging from average patient review scores, referral patterns, affiliated hospital scores, relative costs, and provider awards. Specifically, we used data collected from Centers for Medicare & Medicaid Services (CMS) and provider profile websites on a set of 600,000 US health care providers. We also collected ranking data from other sources; specifically, U.S. News was used for specialty-specific hospital rankings. We converted each provider's information to a set of intuitive qualitative attributes. For instance, affiliated hospitals were mapped to specialty-specific rankings to assign a score to the affiliated hospitals of a provider relevant to their specialty. As a peer-nominated award, we selected the Castle Connolly award. Each year, Castle Connolly distinguishes top providers both nationally and regionally through a peer nomination process that involves over 50,000 providers, and hospitals and health care executives [2]. Similarly, we collected quality data from National Committee for Quality Assurance (NCQA) for each insurance plan ranging from state, plan category, ranking, overall review scores, customer satisfaction scores, as well as preventive care and treatment scores [3].

We then adopted a data-driven approach to determine if the provider attributes were correlated with the insurance quality indicators. Specifically, we measured the correlation between several provider attributes (reviews rating, awards, affiliated hospitals, etc.) of member providers of an insurance plan to key quality scores of the insurance plans. Key challenges to our data collection and analysis included mapping providers from CMS to providers in provider profile sites, mapping insurance names between accepted insurances obtained from provider profile sites and insurances obtained from NCQA, and mapping hospital names between each source. These challenges are due to the lack of a common identifier for providers, insurance plans, or hospitals across the data sources. There have been several studies to determine the quality of health insurance plans. These studies can be split into two categories: (1) health insurance marketplaces and search sites, and (2) attributes associated with health plan quality.

2.1.1 Online Health Insurance Marketplaces and Search Sites

There are several health insurance marketplaces, authorized by the Affordable Care Act, that help individuals and small employers shop for, select, and enroll in high-quality, affordable private health plans. In fact, the Affordable Care Act requires the US

Department of Health & Human Services to develop quality data collection and reporting tools such as a quality rating system, a quality improvement strategy, and an enrollee satisfaction survey system [1]. Information from the quality rating system, quality improvement strategy, and surveys will inform consumer selection of a quality health plan, decisions about quality health plan certification, and the Federal and State marketplaces' monitoring of quality health plan performance. All these measures use data collected through consumer experience surveys such as enrollee experience surveys and health insurance marketplace surveys. Other insurance search sites, such as einsurance.com and insure.com, collect user feedback regarding each interaction with their partner insurance providers. This feedback enables them to identify potential customer service issues and is also used as an essential component of the ranking system that they use to determine how these partners are presented to prospective future clients [4, 5]. Hence, most of these studies focus on user-generated content and do not consider the rich set of provider data readily available. Research is lacking on the association between information from providers in the network with the respective health insurance plans. For example, if patients rate insurance plans based on cost, are these ratings useful for finding providers that provide quality health care?

2.1.2 Attributes Associated with Insurance Quality

Several surveys have examined the quality of health insurance plans based on consumer feedback and have tried to determine attributes associated with insurance quality. Feldman states that a cornerstone of high-quality integrated care for people with medical, behavioral, and long-term services and support needs is a dynamic person- or family-centered plan of care built on significant individual and caregiver involvement and comprehensive assessments and reassessments over time to capture changes in people's circumstances and preferences. Other key ingredients identified were (1) a multidisciplinary care team with one accountable care coordinator, and (2) a comprehensive provider network with a strong primary care base and a range of other providers and services that can accommodate diverse needs throughout a lifetime [6].

URAC (Utilization Review Accreditation Commission), which is an independent, nonprofit organization known for promoting health care quality through its accreditation, education, and measurement programs, addresses the following key areas aimed at helping plans deliver safe, high-quality, patient-centered, high-value care: Wellness and Health Promotion; Care Coordination; Medication Safety and Care Compliance; Rewarding Quality; Care Delivery through a Network; Mental Health Parity; Measures—patient centeredness, coordination of care, patient safety, health plan administration, efficiency, effectiveness of care and health information technology integration; and Patient Experience of Care (Consumer Assessment of Healthcare Providers and Systems Survey) [7]. In our study, we examine the correlation of provider attributes to quality indicators of health insurance plans.

2.2 Methods

2.2.1 Summary

For the purpose of our data-driven analysis, we have collected a large amount of information about US health providers, mainly physicians, from multiple online sources including the CMS data on providers and hospitals, U.S. News rankings of hospitals, and additional provider information and reviews from provider profile websites. We have also collected information about the rankings of private, Medicare, and Medicaid health insurance plans from NCQA. We then mapped entities across sources to create a database of providers and health plans. Figure 1 shows the process of mapping insurances accepted by the providers and the insurance plans obtained from NCQA. We then used this providers' information and insurance information database in each of our analyses.

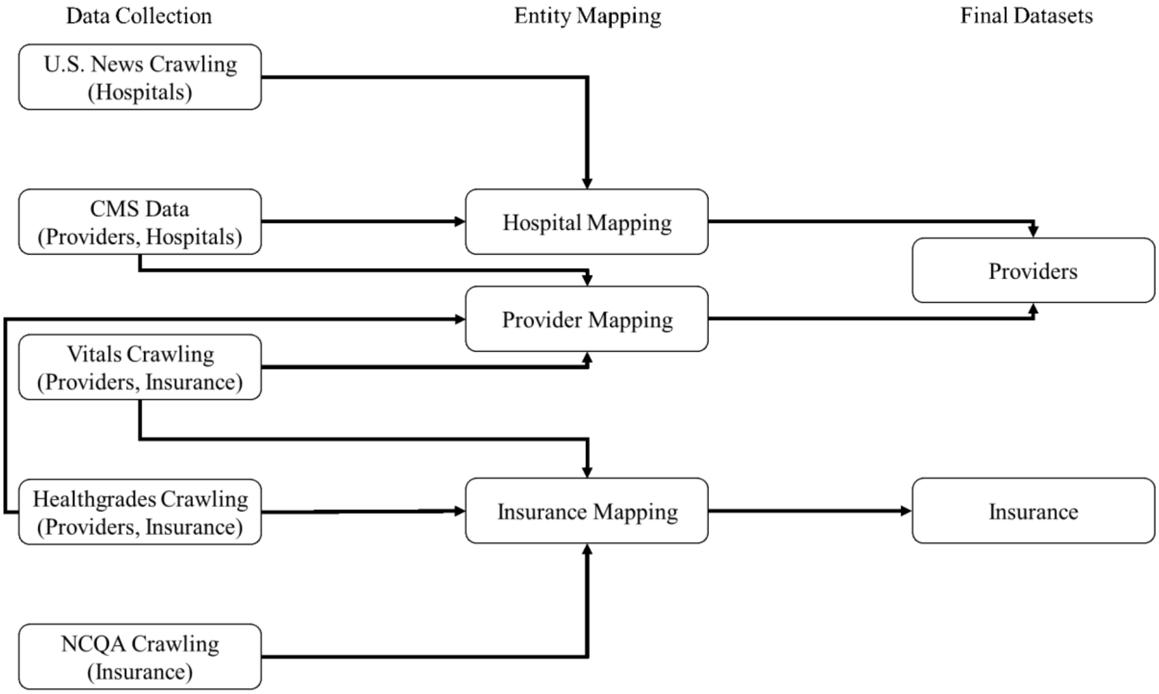

Data Collection
Entity Mapping
Final Datasets
U.S. News Crawling (Hospitals)
CMS Data (Providers, Hospitals)
Vitals Crawling (Providers, Insurance)
Healthgrades Crawling (Providers, Insurance)
NCQA Crawling (Insurance)
Hospital Mapping
Provider Mapping
Insurance Mapping
Providers
Insurance

Figure 1. Visual description of data preprocessing.

2.2.2 Data Collection

Insurance information and patient ratings of providers were collected from both Vitals and Healthgrades [8, 9]. Hospital rankings were collected from U.S. News reports [10, 11]. Additionally, insurance plan rankings for 2014-2015 were collected from NCQA. We also used the datasets released by CMS for health care providers (and hospitals) based in the United States. This information includes general information such as the provider's specialties, medical training, and hospital affiliations [12, 13]. Other provider information includes the Healthcare Common Procedure Coding System (HCPCS), physician referrals, and prescription data [14-16]. Note that all CMS datasets link providers using a National Provider Identifier (NPI). CMS hospital information includes names, location, and a unique identifier, which is used to link each NPI to their affiliated hospitals. CMS data were downloaded directly from CMS websites. Separate crawlers were built using jsoup [17], a Java library for obtaining and parsing webpages, for each of the other data sources: Vitals, Healthgrades, U.S. News, and NCQA.

Aggregating the datasets posed unique challenges for entity mapping, such as mapping providers from Healthgrades to providers in CMS, as described in the next section. In total, we collected information on 3.2 million distinct providers from CMS, 4600 distinct hospitals from CMS, 1.9 million distinct providers from Healthgrades, one million distinct providers from Vitals, and 1956 hospitals from U.S. News. We also collected information of 1264 health plans from NCQA. Of these, NCQA has ranked 1051 plans based on clinical performance, member satisfaction, and results from NCQA Accreditation surveys. The remaining insurances had partial data. After appropriate data transformations and entity mappings, we generated the set of provider attributes listed in Table 1 and health insurance plan attributes listed in Table 2.

Table 1. List of provider attributes used in our analysis based on the data collected.

Category	Attribute	Description	Source	Min.	Max.	Mean	Median
General information	NPI	National Provider Identifier.	CMS	N/A	N/A	N/A	N/A
	Gender	Male or Female, as specified in the CMS data.	CMS	N/A	N/A	N/A	N/A
	Specialties	A set of attributes, one for each specialty, e.g. cardiologist.	CMS	N/A	N/A	N/A	N/A
From peers	NumReferrals	Normalized number of referrals.	CMS	0	4018	70.1	10
	CastleConnolly	Whether or not the provider is recognized by Castle Connolly as a distinguished provider.	Vitals	N/A	N/A	N/A	N/A
Average rating from patient reviews	UserRatings	Overall review score assigned by user (patient).	Reviews from Vitals and Healthgrades	0	100	82.06	87.5
	NumReviews	Number of patient reviews for the provider.	N/A	0	247	0.96	0
Insurance	NumInsurances	Number of insurers accepted by the provider.	Vitals and Healthgrades	1	8	1.7	1
	IndividualInsurers	A set of attributes, one for each insurer accepted by the provider, e.g. Humana.	Vitals and Healthgrades	N/A	N/A	N/A	N/A
Hospital affiliations	HospitalRanking	The ranking of the provider's affiliated hospitals.	CMS (hospitals) and U.S. News (ranks of hospitals)	N/A	N/A	N/A	N/A

Table 2. List of health insurance attributes used in our analysis based on the data allocated[1].

Category	Attribute	Description
General information	PlanName	Insurance plan name.
	State	The state to which the plan belongs.
	PlanCategory	The category of the plan, e.g. private, Medicare, Medicaid.
	PlanType	The type of the plan, e.g. preferred provider organization (PPO), health maintenance organization (HMO).
Quality indicators – Overall	Rank	The overall rank of the plan.
	OverallScore	The overall score of the plan.
Quality indicators – Customer service	OverallConsumerSatisfactionScore	The score for consumer satisfaction.
	GettingCareScore	Scores based on appointments, preventive care, test, and easy and quick access to treatments.
	SatisfactionWithPhysiciansScore	Scores based on providers, care revived and health promotion and education.
	SatisfactionWithHealthPlanServicesScore	Scores based on handling claims and other plans services.
Quality indicators – Prevention	OverallPreventionScore	The score for preventive care.
	ChildrenAndAdolescentsScore	Scores based on well-child visits, immunizations, nutrition counseling, physical activity counseling.
	Women'sReproductiveHealthScore	Scores based on prenatal checkup and postpartum care.
	CancerScreeningScore	Scores based on various cancer screenings.
	OtherPreventiveServicesScore	Scores based on flu vaccinations, chlamydia screening, and other preventive care.
Quality indicators – Treatment	OverallTreatmentScore	The score for different treatments.
	AsthmaTreatmentScore	Scores based on asthma medication and treatment.
	DiabetesTreatmentScore	Scores based on blood pressure control, glucose testing and control, low-density lipoprotein cholesterol screening and control, monitoring kidney diseases.
	HeartDiseaseTreatmentScore	Scores based on controlling blood pressure and cholesterol and beta-blockers after heart attack.
	MentalAndBehavioralHealthScore	Scores based on depression medication, alcohol and drug dependence treatment, etc.
	OtherTreatmentMeasuresScore	Scores based on monitoring key long-term medications, antibiotic use, testing for chronic obstructive pulmonary disease, etc.

[1] All attributes in this table are from NCQA.

2.2.3 Entity Mappings

The names of insurance obtained from Vitals and Healthgrades differ from the names of insurance in the NCQA data. For example, “United Healthcare Services, CA” and “United Healthcare, CA” refer to the same insurance plan, as do “Aetna Life Insurance, AR” and “Aetna HMO, AR”. In order to achieve this mapping, we used the Levenshtein distance metric [18] to map Healthgrades and Vitals insurance to NCQA insurance. This generated 242 mappings between Vitals and NCQA insurance and 1330 mappings between Healthgrades and NCQA insurance.

The hospital rankings listed by U.S. News categorize hospitals across several specialties for adults and children; for each hospital listed, the hospital’s score, name, and location were collected for each specialty for both adults and children. Further, the hospital specialties reported by U.S. News do not always correspond to the specialties listed by CMS. In particular, CMS uses a taxonomy of medical specialties that consider subspecialties, whereas U.S. News uses broad categories of specialties [19]. Note that this mapping is not necessarily one-to-one; for example, a provider specializing in internal medicine may map to several categories listed by U.S. News. Therefore, we manually mapped all specialties with more than 100 occurrences to the specialties used by U.S. News. This generated 5651 mappings. We then used these mappings to assign scores to each of the affiliated hospitals, using the average for a hospital’s score when the provider’s specialty mapped to more than one specialty listed by U.S. News. We then assigned HospitalScore to the hospital affiliation with the maximum score, where null values are used for providers whose hospital affiliations are missing from the mappings.

Also, for each HCPCS code of a provider, we computed the amount charged for this provider, relative to others of same specialty in the area (1000 closest within a 30-mile radius, normalized to a range of 0 to 100, where 100 goes to the most expensive physician). We then took the weighted average (by the number of procedures of a provider) of these relative charges to get the relative cost with respect to area.

In order to identify Castle Connolly and patient reviews information for each provider, CMS providers needed to be mapped to Vitals and Healthgrades provider profiles. This mapping exercise allowed us to map 608,935 providers between CMS, Vitals, and Healthgrades, 25,514 of whom have received a Castle Connolly award. To map CMS providers to providers in the other sources (Heathgrades and Vitals), we followed a hybrid automatic-manual data integration approach. First, we identified a promising set of attributes to use for mapping, specifically, first name, middle name, last name, address, medical school, graduation year, affiliated hospitals, and specialties. For each attribute, we constructed a customized mapping algorithm. For example, the mapping between first names is computed using the Levenshtein distance between the two strings. Then, we assigned weights to each attribute matching score based on a large number of accuracy experiments, where the authors defined the ground truth mappings. We then computed a mapping threshold based on the mapping scores via more accuracy experiments. Note that each Vitals/Healthgrades provider is mapped to at most one CMS provider, so no duplicate provider data are present in the final dataset.

Only 4% of all mapped providers have received a Castle Connolly award, and 42% of all mapped providers have zero referrals. A majority of providers with zero referrals specialized in Internal Medicine, Family Medicine, or Emergency Medicine. Also, 213 of 1264 health plans collected had incomplete data. In order to correlate rank of affiliated hospitals and insurance scores, we needed the rank of the hospitals. However, only 50 out of the 1956 hospitals obtained from U.S. News were ranked. We considered the unranked hospitals to be at the bottom of the list. We then took the median of the unranked hospitals (i.e. 1053) and considered this to be the rank of the unranked hospitals. Also, in order to account for local trends, we performed our analysis at both the national and state levels. Health care is regulated at both the state and federal levels. These regulations, along with demographics and population health, create localized trends in health care.

2.3 Results

2.3.1 Summary

The results of our analysis consist of a description of general statistics about the different types of insurance and a state-wise analysis of the consumer satisfaction insurance plans. Then we report on correlations between insurances' consumer satisfaction score and the average patient review scores of providers that accept those insurances. We report similar correlations between insurances' overall NCQA consumer satisfaction score and then average number of referrals per provider, ratio of Castle Connolly providers, average affiliated hospital scores of providers, and relative cost of providers with respect to area. Last, we break down the providers according to their specialties and describe correlations between the average patient review scores and treatment insurance scores for condition-specialty combinations.

2.3.2 General Statistics of Insurance Plans

We first analyzed general statistics about the various insurance plans at the national level. We calculated the average overall consumer satisfaction scores of the insurance plans (see corresponding row in Table 2), where we average across the types of insurance plans: private, Medicare, and Medicaid. We also calculated the average patient review scores of providers (referred as "UserRatings" in Table 1) accepting these different types of insurances. Our findings are shown in Table 3 along with the statistical analysis. The patient review scores are on average higher than the insurance satisfaction scores, and with high significance for private PPOs and Medicare plans.

Table 3. General statistics about different types of health insurance plans.

Insurance plan type	**Average patient review score (*p* value)**	**Average consumer satisfaction insurance score (*p* value)**
Private PPO	82.03 (< 0.001)	79.75 (0.384)
Private HMO	82.54 (< 0.001)	81.63 (< 0.001)
Medicaid	82.78 (< 0.001)	77.52 (< 0.001)
Medicare PPO	82.39 (< 0.001)	76.71 (0.263)
Medicare HMO	81.55 (< 0.001)	76.9 (0.123)

To estimate significance between values in the same row of Table 3, the Wilcoxon signed-rank test significance values are as follows, between average patient and insurance scores: private PPO < 0.001, private HMO = 0.13, Medicaid = 0.008, Medicare PPO < 0.001, and Medicare HMO < 0.001. To compute significance of a value with respect to the union of the other four plan types in the same column (*P* value), we used the Mann-Whitney U test.

We also computed the average consumer satisfaction insurance scores for each state. The heat map in Figure 2 shows our findings. The darker colored states are those that have a higher overall consumer satisfaction insurance score while the lighter ones have lower consumer satisfaction insurance scores. From the map, we can conclude that northeastern states have higher consumer satisfaction insurance scores.

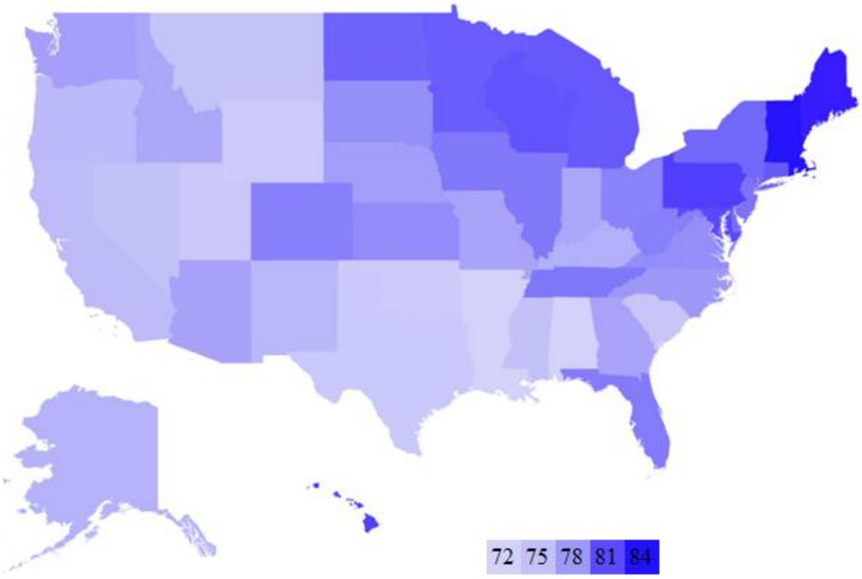


Figure 2. Heat map showing average consumer satisfaction insurance scores of different plans.

Similarly, we computed the number of health care providers per 1000 people for each state. As shown in Figure 3, the darker colored states have more providers per capita while the lighter states have fewer per capita. From this map, we can see that the northeastern states also tend to have more health care providers per capita.

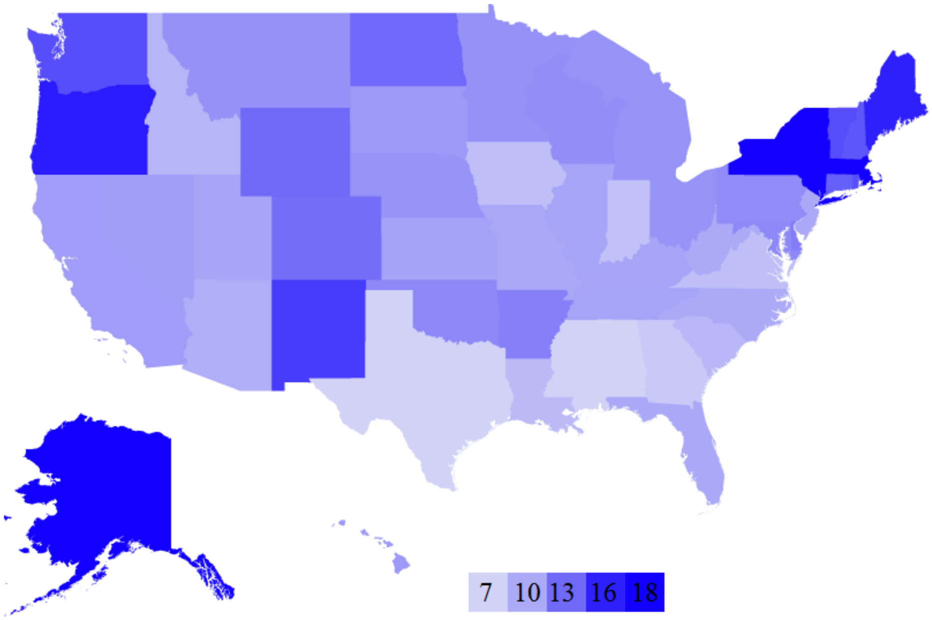


Figure 3. Heat map showing number of health care providers per 1000 people in each state.

Finally, we counted the number of insurance plans evaluated by NCQA per state. The heat map in Figure 4 shows our results. The darker colored states have more insurance plans while the lighter ones have fewer. The map shows that the most populous states have the most insurance plan options while the less populous states tend to have fewer.

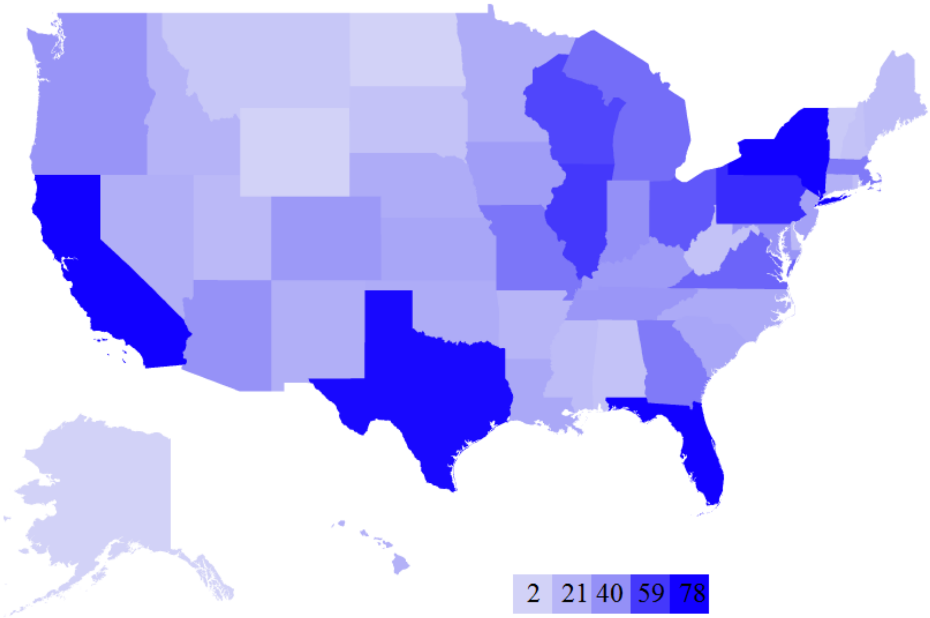


Figure 4. Heat map showing the number of health insurance plans evaluated by NCQA per state.

2.3.3 Attribute Correlations

We computed the Pearson correlation of average patient review scores of providers that accept a particular insurance plan and that insurance plan’s NCQA scores. We found that there is a moderate positive correlation between these attributes (specifically 0.376). Figure 5 illustrates this correlation. We then did the same analysis state-wise and found that the Pearson coefficient increases in value, showing greater correlation when we localize the analysis. Table 4 shows the correlation coefficient between these same attributes for some of the different states. A couple of interesting observations can be made based on these correlations. First, there seems to be a moderate correlation between average patient review scores and consumer satisfaction insurance scores. Hence, insurance that includes providers with good reviews is more likely to have a better overall score. Also, the correlation between these two attributes seems to get stronger when we break down the data state-wise.

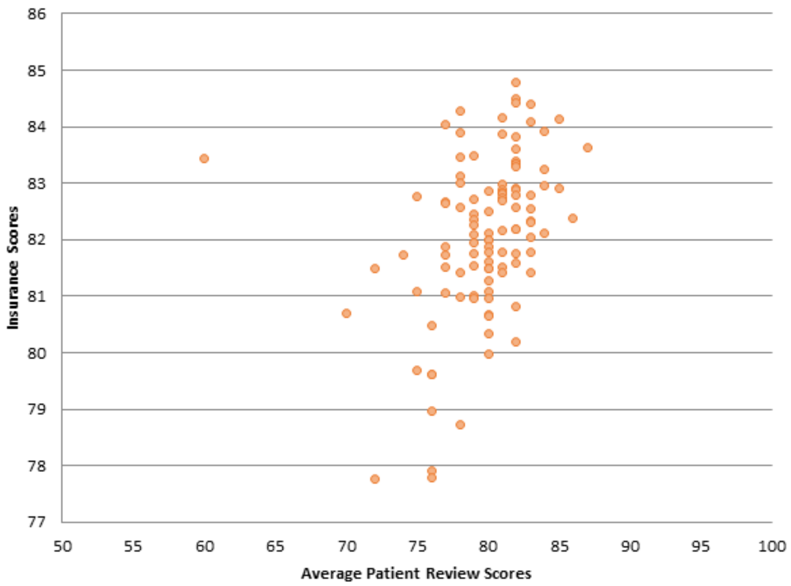


Figure 5. Correlation between average patient review scores and consumer satisfaction insurance scores (overall)[2].

Table 4. Correlation between average patient review scores and consumer satisfaction insurance scores.

State	**Correlation**
Overall	.376
State-wise	
New York	.869
Texas	.794
Illinois	.738
Pennsylvania	.696
California	.647
Ohio	.549
Florida	.457

Next, we report correlations between average referrals per provider for insurances and those insurances' NCQA scores. Our analysis showed that there is a positive but very low correlation (specifically 0.031) between these two attributes. Hence, referral frequency of providers is negligibly correlated to consumer satisfaction insurance scores. Figure 6 further illustrates this correlation. Figure 7 illustrates the correlation between ratios of

[2] Correlation coefficient = 0.376, $p < 0.001$.

providers having the Castle Connolly award to the overall insurances' NCQA scores. We found a positive but negligible relationship between these attributes, specifically 0.183. Hence, whether a provider has received a Castle Connolly award or not does not affect the insurances' overall score. With respect to correlation between average ranks of affiliated hospitals and consumer satisfaction insurance scores, there exists a negative but negligible correlation between these two attributes (specifically -.108). Since we are considering ranks of hospitals, the negative correlation is expected. Hence, consumer satisfaction insurance scores are unlikely to be affected by the ranks of affiliated hospitals of the providers under that insurance plan. Figure 8 illustrates this correlation. We also determined the correlation relationship between relative cost of providers with respect to area and the consumer satisfaction insurance scores. Our findings showed a weak positive correlation of 0.266 between these two attributes. Figure 9 shows this correlation.

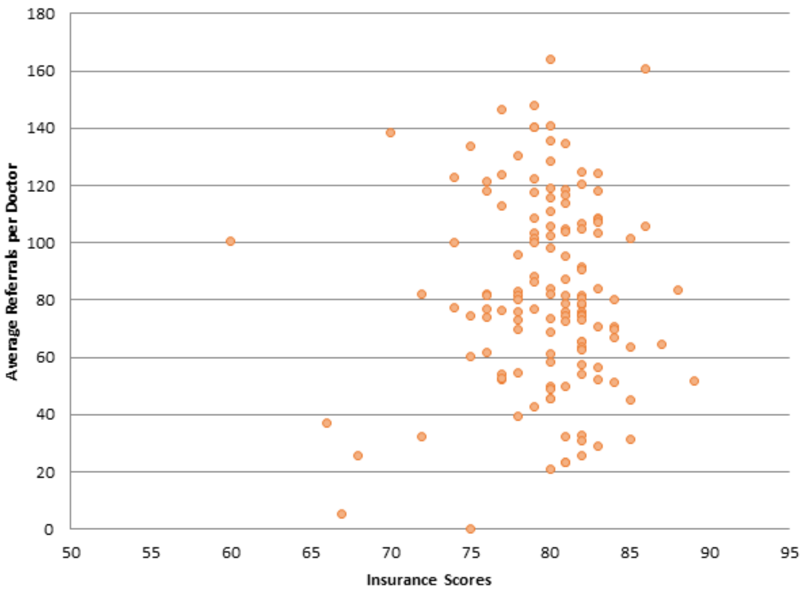


Figure 6. Correlation between average referrals per provider and consumer satisfaction insurance scores[3].

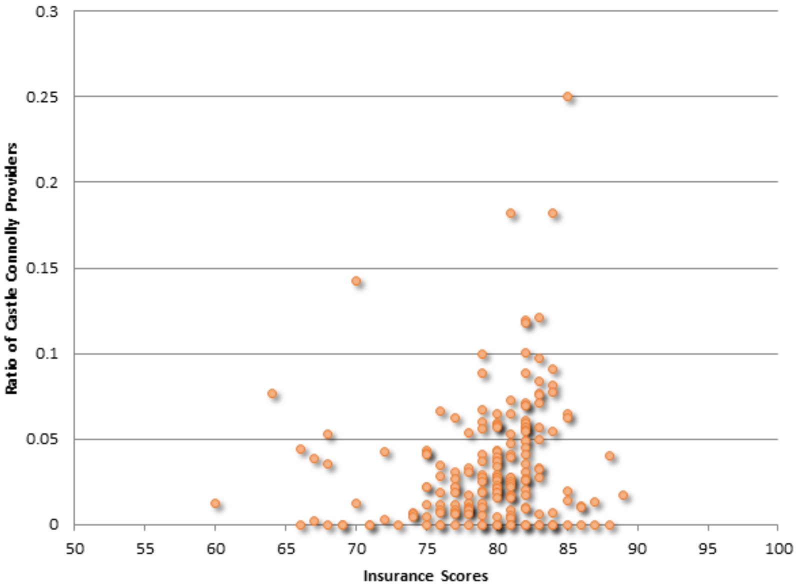


Figure 7. Correlation between ratio of Castle Connolly providers and consumer satisfaction insurance scores[4].

[3] Correlation coefficient = 0.031, $p = 0.715$.
[4] Correlation coefficient = 0.183, $p = 0.001$.

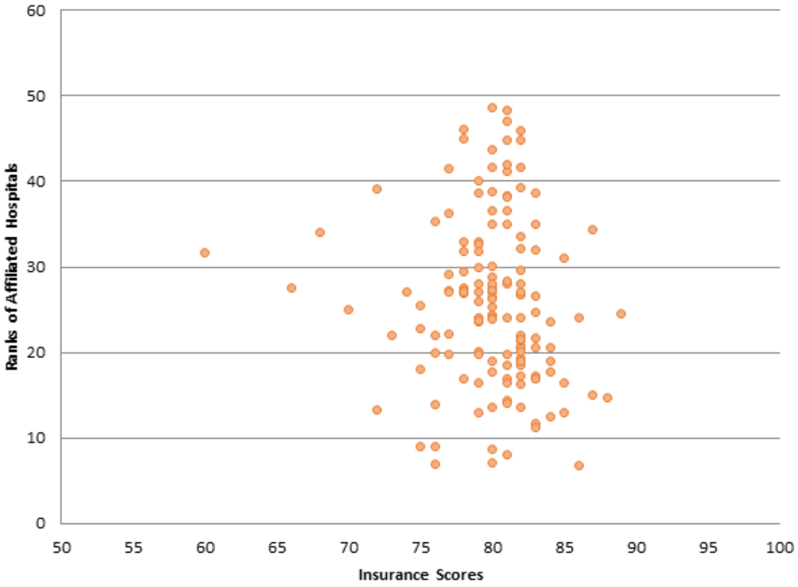


Figure 8. Correlation between ranks of affiliated hospitals and consumer satisfaction insurance scores[5].

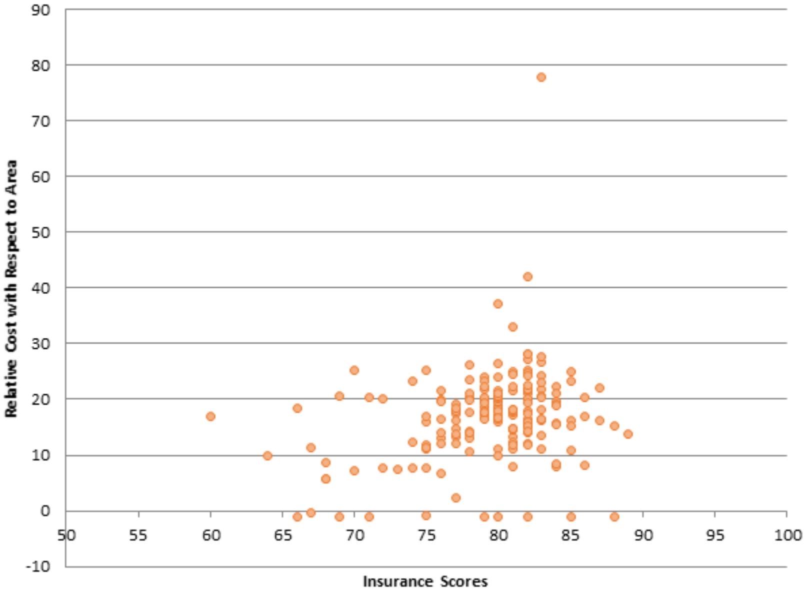


Figure 9. Correlation between relative cost of providers with respect to area and consumer satisfaction insurance scores[6].

[5] Correlation coefficient = 0.108, $p = 0.199$.

[6] Correlation coefficient = 0.266, $p < 0.001$.

We then examined correlations between average patient review scores for specialist providers and the NCQA treatment insurance scores for these specialties. For this we used the individual treatment scores obtained from NCQA for the various conditions described in Table 2. We then compared these scores to the average patient review scores of only those providers that provide that kind of care, as shown by the mapping of condition to specialties in Table 5. For example, the average patient review scores of pediatricians were compared to the NCQA scores for treatment of children and adolescents. Table 5 lists our findings. We observed that for women's health, mental and behavioral health, and cancer screening there exists a positive but negligible correlation between the average NCQA scores and the average patient review scores. However, for heart diseases, child and adolescent health, and diabetes, there exists a negative and negligible to weak correlation between the attributes.

Table 5. Conditions and associated specialties ranked by correlation between NCQA scores and average patient review scores.

Condition from NCQA	**Corresponding member specialties**	**Correlation of treatment insurance score with average patient review score**
Women's health	Obstetrics and Gynecology, Gynecology Oncology	0.135
Mental and behavioral health	Counselor, Psychoanalyst, Clinical Neuropsychologist, Psychologist, Psychoanalysis, Marriage and Family Therapist	0.112
Cancer screening	Pediatric Oncology, Oncology, Hematology & Oncology, Radiation Oncology	0.112
Heart disease	Cardiologist, Cardiac Rehabilitation, Cardiology Technician, Cardiovascular Diseases	-0.002
Children and adolescent health	Pediatrics, Neonatal Pediatrics, Pediatrics Critical Care	-0.083
Diabetes	Diabetes Educator, Endocrinology, Diabetes and Metabolism	-0.259

2.4 Discussion

2.4.1 Principal Findings

Our analysis shows that there are several provider attributes that are correlated to insurance quality attributes. We showed that patient review scores for providers are correlated to consumer satisfaction insurance scores. This is expected given that patients who are happy with the care they receive from their providers are more likely to also be happy with their overall insurance plan. For example, if a patient has complaints about the billing at a provider's office, this patient will likely be unhappy with the insurance company who did not help cover or settle the bill.

On the other hand, our results showed negligible correlation between average referrals per provider and consumer satisfaction insurance scores. This is not surprising, as there is no convincing evidence that a higher number of referrals is connected to better skills for a provider or to better relationship with patients. Similarly, we demonstrated that there is a negligible correlation between the ratio of Castle Connolly providers and the consumer satisfaction insurance scores.

The case between rank of affiliated hospitals and consumer satisfaction insurance scores was similar. However, we found a weak positive correlation between the relative cost of providers with respect to their geographic area and consumer satisfaction insurance scores. This may be explained by the fact that providers with satisfied patients may increase their prices. Of course, the charged prices are not so important, as Medicare and Medicaid generally have fixed compensations per procedure.

Our results on the lack of correlation of patient reviews score and treatment quality metrics for various conditions may indicate that patients who are satisfied with their provider may not necessarily have better health outcomes, as studies have shown that patients often rate their providers based on non–outcome-related attributes such as wait and visit times. For instance, research has shown that the average satisfaction score for wait times of 0-15 minutes was 94.3 on a 100-point scale [20].

Our findings can be used to help consumers make informed choices about their insurance plans. Health insurance marketplaces may find patient review scores for providers of each insurance plan to be a useful addition to other insurance plan metrics. Alternatively, consumers can use this information in their own research to identify potential insurance plans based on the review scores of providers on review sites such as Vitals and Healthgrades.

Further, insurers may use our results to better understand the relationship between their patients' satisfaction and their network of providers. For example, although it is not clear if there is a cause-effect relationship, our results indicate that hiring a provider with high patient review scores may contribute more to the overall consumer satisfaction insurance plan rating than hiring a provider who has been receiving many referrals from their colleagues. Further, our results indicate that more expensive providers are correlated with higher plan satisfaction, which seems to be at odds with the providers' "tier-ing" approach of insurers, who try to encourage patients to visit the cheaper providers.

Health care providers may also use our results to decide which insurance plans to accept. As noted above, a patient whose bill was not covered by an insurance company may complain about the billing at the provider's office on a provider review site, leading to a lower overall patient review score. A provider wishing to maintain a favorable score may thus choose to avoid accepting insurance plans with low consumer satisfaction scores.

2.4.2 Limitations

One of our biggest limitations is that not all of the data we obtained are complete. For example, a majority of the providers have zero reviews; this is likely due to the fact that only 4% of Internet users post online reviews for providers, and previous work has shown that most providers have zero reviews [21]. Similarly, a majority of the hospitals had no ranking information. A second limitation is that we sourced our data from multiple sites such as Vitals, CMS, Healthgrades, and NCQA. We then tried to map the various attributes across these sources. However, the accuracy of these data sources cannot be guaranteed. Another limitation is that referral frequency is greatly influenced by the specialty of the provider, and hence it needs to be normalized in terms of specialty in order to be used as an effective quality measure. Also, while the Castle Connolly award is prestigious and rigorously vetted, the award is biased towards providers who have more experience.

2.4.3 Conclusions

Our data-driven analysis led to several interesting findings. Higher consumer satisfaction insurance scores are correlated with their providers having better patient review scores. There also seems to be a correlation between cost of medical care and insurance ratings. However, there was negligible correlation between other quantitative attributes such as number of referrals per provider, ratio of Castle Connolly award recipients, affiliated hospitals scores, and health insurance ratings. These findings may provide new insights into what attributes should be adopted by insurance marketplaces and search portals to empower patients in a patient-centered setting.

Chapter 3

The Impact of Colleges and Hospitals to Local Real Estate Markets

This paper studies how the presence of universities and hospitals influences local home prices and rents. We analyze the data on ZIP code level and on the level of individual homes. Our ZIP code-level analysis uses median home price data from 13,105 ZIP codes over 21 years and rent data from 15,918 ZIP codes over 7 years to compare a ZIP code's appreciation, volatility and vacancies to the size of a university or hospital within that ZIP code. Our home-level analysis uses data from 2,786,895 homes for sale and 267,486 homes for rent to study the impact of the distance from the nearest university or hospital to individual home prices. While our results generally agree with our expectations that larger, closer institutions yield higher prices, we also find some interesting results that challenge these expectations, such as positive correlations between volatility and university/hospital size in some ZIP codes, a positive correlation between rent and distance from a hospital for some homes, and lower correlations of rent vs. distance from a university compared to price vs. distance.

3.1 Introduction

Home price is made of two parts: price of land and the cost of the house. Land value is derived from its location which often, especially in urban areas, accounts for the lion's share of overall home price. The value of the land is subject to the laws of supply and demand and in turn depends on the land's scarcity. Indeed, decoupling price of land from price of construction has been extensively researched [22, 23]. Many factors are baked into land price, including proximity to amenities and land's inherent quality (e.g., proximity to a shoreline, the mountains, etc.). Unique and somewhat subjective home characteristics like view as well as proximity to ocean, lake, etc. are known to influence home price [24]. Conversely, land price may be adversely affected by proximity to sources of noise and pollution (airports, major highways, etc.) [25, 26]. Unlike building material, labor and capital, land is a "finite," or "non-renewable," resource, often limited by stringent geographic and topographical constraints. Amenities pertain to proximity and accessibility to things like opportunities in employment, education, transportation, entertainment, retail, cultural, recreational, etc.

This analysis focuses on universities and hospitals as "opportunity hubs," which encapsulate "packaged amenities" in terms of those listed above. It studies the impact of these institutions on both home sale price and rent. Both types of institutions attract a "stable," educated and mobile workforce, a mix of demographics and incomes, and various amenities. Unpacking amenities isn't altogether simple, and is a somewhat subjective art. For example, a neighborhood's school rating is on one hand a reflection of the neighborhood and its characteristics, demographics and economics. Conversely, the rating of a neighborhood's schools affects its home prices, primarily via the value of the land on which each home in the neighborhood is built. Throughout this paper, we will unravel pricing substructure via the correlations between home value and proximity to said institutions as well as their "idiosyncratic rhythm."

This study focuses on US homes. We perform two types of analysis. In *ZIP code-level analysis*, we use the median home price per ZIP code and study how the containment of a university or a hospital affects (correlates to) home prices by comparing against ZIP codes that do not contain such institutions. As real estate is always local, we looked for a wide availability of data at a very local possible level and ZIP code level data fit our requirements [27]. In *home-level analysis*, we consider the prices of individual homes with respect to their exact distance from institutions. The goal of this analysis is to study how the distance of a property to an institution affects its price or rent and test our assumption that universities and hospitals generally increase the sale prices and rent of nearby homes. Our basis for this assumption includes both prior work that has analyzed the effect of real estate prices in relation to proximity of various features as described above and in the Related Work section as well as intuition; for example, one would expect that homes close to universities have higher rent as students without cars prefer them. In addition to computing the price and rent correlations with the distance from a university or hospital, we study how rent or price appreciation and volatility as well as vacancies change over time and how they correlate with the size of a university or hospital.

We collected data for home listings and historic home rent and price trends from public listing sources. We also collected ZIP code populations, hospitals and vacancies data from government sources. We built a university dataset by crawling and combining data from online sources (US News and Wikipedia).

Our results show several correlations. In our ZIP code-level analysis, we found the strongest correlations in ZIP codes with a population below the national ZIP code average. These correlations were between appreciation and hospital size, volatility and university size, volatility and hospital size, and vacancies and university size. In our home-level analysis we found significant correlations between rent and distance from a university (especially private universities), and rent and distance from a smaller hospital.

3.2 Related Work

Real estate prices have been a frequent subject of analysis. Cesa-Bianchi et al. compared house prices in advanced and emerging economies between 1990 and 2012 and found that house prices in emerging markets experience faster growth, more volatility, less persistence and less synchronized than house prices in advanced economies [28]. Favara and Imbs found that housing prices increased in response to the expansion of mortgage credit [29]. Muehlenbachs et al. found that shale gas development has a negative effect on property values in areas dependent on groundwater, but a positive effect on property values in areas with piped water [30]. Waddell et al. drew several conclusions from their analysis of residential property values in Dallas County, Texas: including a significant but fairly localized central business district price gradient; improvements to modeling the price effect of proximity to employment centers and other nodes of activity; amenities such as highways, retail, universities, and hospitals had a significant effect on modeling housing values; and a significant influence of race on housing prices [31]. Nau and Bishai found that life expectancy within communities predicted increases in home price indexes [32]. Otto and Schmid analyzed real estate prices in Germany using spatiotemporal models and found that urban regions with higher population density and higher per-capita disposable income have higher land prices than rural areas, shocks in regional real estate prices “ripple out” and affect the whole economy, and population density had an increasing impact on real estate prices [33].

Several papers have evaluated the impact of nearby points of interest on home prices. Rascoff and Humphries found that homes within a quarter mile of Starbucks locations appreciated more quickly than the overall rate of nationwide home appreciation [34]. Turner found that a several of points of interest, including supermarkets, restaurants and movie theaters, increase nearby home values in three neighborhoods in the San Francisco Bay Area [35]. Bolitzer and Netusil found that open spaces such as parks and golf courses increased nearby home prices in Portland, Oregon [36]. Similarly, Anderson and West’s analysis of the Minneapolis-St. Paul metropolitan area found that open spaces provide more value to homes in certain neighborhoods, such as those near the central business district or with many children [37]. Debrezion et al. found that real estate prices in three Dutch metropolitan areas are affected more by the most frequently chosen railway station in an area than the nearest station, and this effect is more pronounced in more urbanized areas [38].

Other work has analyzed economic statistics in populations near universities and hospitals. Moore and Sufrin concluded that large nonprofit institutions such as universities and hospitals can generate employment and personal income through interregional trade [39]. Beeson and Montgomery found that employment growth rates and income are higher in areas with higher-ranked universities, the probability of being employed as a scientist or engineer increases with local university research and development funding, and the probability of being employed in a high-tech industry increases with the number of local university graduates [40]. Hedrick et al. found that university commercial activities reduce private employment in small counties, particularly in the areas of finance, insurance, and real estate, but university enrollment and spending increase local employment, leading in a net positive effect on employment [41]. Moore's analysis of the State University of New York university system found that per capita income generation in counties with a university is negatively correlated with per capita personal income; in other words, the greatest impact on income generation per capita is found in counties with lower personal incomes [42].

3.3 Methods

3.3.1 Data Collection

Median ZIP Code Home Price and Rent

Zillow maintains a dataset of home and rental data for public use [43]. For our ZIP code home price analysis, we used the ZIP Code Zillow Home Value Index (ZHVI) data for May 2017, which lists the median home price in 13,105 ZIP codes for each month from no earlier than April 1996 to May 2017. For our ZIP code rent analysis, we used the May 2017 ZIP Code Zillow Rent Index (ZRI) data, which lists the median rent in 15,918 ZIP codes for each month from no earlier than November 2010 to May 2017. Apartments were not included in the ZRI calculation, thus our statements regarding rent in our ZIP code-level analysis refer only to the rent of houses. In our ZIP code-level analysis, we use the terms “home price” and “rent” to refer to ZHVI and ZRI, respectively. Note that these amounts are computed based on Zillow’s estimate of market price and rent.

ZIP Code Population Data

For population data, we used the 2010 census data provided by the United States Census Bureau [44]. We used two datasets extracted from this data for our analysis: a list of ZIP code tabulation area (ZCTA) populations, and a list of ZCTA population densities, where population density is given by the average number of people per square mile. We assume that the population and population density of each ZCTA are equal to the population and population density of the ZCTA’s corresponding ZIP code.

Universities

We collected university details via a twofold approach restricted to universities in the United States. The first step consists of collecting details about the universities in the United States from Wikipedia [45]. This data source provides many crucial details about the university such as name, number of enrolled students, location and university-type to name a few. The second step includes finding rankings for these universities, the data for which is collected from US News and World Report’s ranking and is restricted to the first 200 ranked universities, while the others are unranked [46]. For our ZIP code-level analyses of price, rent and vacancies over time, we use four subsets of ZIP codes based on the number of students enrolled in a university in each ZIP code as described in Table 6. This distribution was selected to give each subset relatively similar sizes between ZIP codes with home price data and ZIP codes with rent data. Each ZIP code that contains more than one university is assigned to the subset corresponding to the university in that ZIP code with the most enrolled students.

Table 6. Distribution of ZIP codes with home price data based on the number of enrolled students.

Number of enrolled students	ZIP codes with home price data	ZIP codes with rent data
0 (No university)	12,473	15,153
Fewer than 10,000 Students	501	611
10,000–20,000 students	73	85
20,000 or more students	58	69

Hospitals

The Centers for Medicare and Medicaid Services (CMS) provide data used by the Medicare.gov website, including data on hospitals and physicians [47, 48]. Using the hospital data, we determined which ZIP codes contained a hospital. To determine the number of doctors each hospital has, we used the affiliated hospitals listed for each physician in the physician data. For our ZIP code-level analyses of price, rent and vacancies over time, we use four subsets of ZIP codes based on the number of doctors affiliated with a hospital in each ZIP code as described in Table 7. As above, this distribution was selected to give each subset relatively similar sizes between ZIP codes with home price data and ZIP codes with rent data. Each ZIP code that contains more than one hospital is assigned to the subset corresponding to the hospital in that ZIP code with the most affiliated doctors.

Table 7. Distribution of ZIP codes with home price data based on the number of affiliated doctors.

Number of affiliated doctors	ZIP codes with home price data	ZIP codes with rent data
0 (No hospital)	10,819	13,009
Fewer than 100 doctors	309	539
100–500 doctors	1496	1837
500 or more doctors	481	533

Home Listings

We collected data related to homes available for rent and sale from an online listings source that provides various details related to each home such as rent/sale price, home address, number of bedrooms and bathrooms, ZIP code and exact location (latitude and longitude). Apartments account for 7% of the rental data. We used each home's latitude and longitude to calculate the distance from any universities in the same ZIP code or neighboring ZIP codes. We cleaned the data, which includes removing entries with no details about the location, rent/sale price and number of bedrooms. In our home price analysis, we use the term "home price" to refer to the listed sale price and "rent" to refer to the listed rent price. A quantitative summary of the data of homes for rent and for sale is shown in Table 8.

Table 8. Home Listings details: Number of records, start and end date of record collection.

Home data type	Number of records	Start date	End date
For rent	267,486	2017-04-03	2017-04-15
For sale	2,786,895	2017-05-03	2017-05-26

Home Vacancy

The US Department of Housing and Urban Development (HUD) provides home vacancy data [49]. This dataset includes the vacancy statistics for homes and businesses within each census tract. Census Tracts are "small, relatively permanent statistical subdivisions of a county or equivalent entity that are updated by local participants prior to each decennial census as part of the Census Bureau's Participant Statistical Areas Program" [50]. We mapped the census tract data to ZIP codes by using the Tract-ZIP code mapping provided by HUD and assuming a uniform distribution of vacant homes in each tract. The vacancy details include statistics such as the count of vacant homes, count of homes, and periods of vacancy.

Table 9 shows a summary of the types of data used and their use in either or both our ZIP code-level analysis and home-level analysis.

Table 9. Data source description and usage summary.

Type	Description	Usage
Median ZIP code home price and rent	Monthly median rent and sale prices by ZIP code	ZIP code-level analysis
ZIP code population data	Demographic data of each ZIP code in USA	Both
University details	Statistical and locational details of 1991 universities	Home-level analysis
Hospital details	Statistical and locational details of 4691 hospitals	Both
Vacancy details	Quarterly statistical details of home vacancies in various ZIP codes in USA	ZIP code-level analysis
Home details	Details of homes for rent and sale throughout USA	Home-level analysis

3.3.2 ZIP Code-Level Analysis

We used two metrics to analyze median home price and rent in each ZIP code. The first of these is average annual appreciation, which is the average difference in median home price or rent in a ZIP code compared to twelve months prior. To calculate this, we sampled the median home price and rent for May of each year. The second metric is volatility. Given P_z, a list of median home price or rent over time in ZIP code z, we define volatility as σ/μ, where σ is the standard deviation of the values in P_z and μ is the mean of the values in P_z.

We also analyzed the percentage of vacancies in each ZIP code. For our analysis, we averaged the ratio of vacant homes over the four most recent quarters in our data (Q3 2016, Q4 2016, Q1 2017 and Q2 2017) for each ZIP code. This was done to account for changes in the vacancy ratio over the course of a year due to homes with seasonal vacancies (e.g. vacation homes).

For ZIP codes that contain a university or a hospital, we analyzed each of these metrics as a function of the size of the university or hospital to observe their correlations. We define size as the number of students enrolled in a university or the number of doctors affiliated with a hospital. We calculate these correlations using the Pearson correlation coefficient. For random variables X and Y, the Pearson correlation coefficient is defined as $\rho_{X,Y} = \mathrm{cov}(X,Y)/\sigma_X\sigma_Y$, where $\mathrm{cov}(X,Y)$ is the covariance of X and Y, σ_X is the standard deviation of X and σ_Y is the standard deviation of Y [51].

In addition to analyzing all ZIP codes together, we also partitioned the ZIP codes across various dimensions and analyzed each partition separately. Table 10 shows these dimensions and the threshold used to split the ZIP codes into two partitions. We also analyzed subsets of ZIP codes in metropolitan areas or non-metropolitan areas.

Table 10. Dimensions and thresholds to partition the ZIP codes into two sets.

Splitting dimension	**Splitting threshold**
Number of students (ZIP codes with universities)	20,000
Number of doctors (ZIP codes with hospitals)	1000
Median ZIP code home price/rent	National ZIP code average
ZIP code population	National ZIP code average
ZIP code population density	National ZIP code average
Metropolitan area ZIP code, as listed in Zillow data	Yes/no

3.3.3 Home-Level Analysis

The home level analysis focuses on the impact of distance to a university or hospital on the home price or rent. This impact is gauged from the correlation of the home price or rent with the distance of the home from the nearest university or hospital. As in ZIP code-level analysis, we explore the Pearson correlations for various subsets of the homes, defined across various dimensions, such as the number of bedrooms or population of their ZIP code, university types and number of doctors in hospitals.

To partition the home data across such dimensions, a key step is to join the university (or hospital) and home data, as described in Table 9, based on the nearest university (or hospital) decided by the home-university (or hospital) distance. Specifically, for each home, we store its closest university (or hospital) if there is more than one institution within the home's vicinity. The result is a pool of homes which are within a defined vicinity range from a university (or hospital). We create separate data pools for home rent and sale price data. Here the defined maximum vicinity is ten miles from the location of the university or hospital. In the analysis we also consider reducing the vicinity ranges, to see if the correlation is stronger if we focus on homes that are very close to the institutions.

Note that the research and analysis did not use any data from HomeUnion.

3.4 Results

3.4.1 ZIP Code-Level Analysis

Home Price and Rent over Time

We grouped ZIP codes into four subsets: ZIP codes with no university, ZIP codes that have a small university (fewer than 10,000 students enrolled), ZIP codes that have a medium university (at least 10,000 but fewer than 20,000 students enrolled) and ZIP codes with a large university (20,000 or more students enrolled). We then compared the average of the median home price and rent over time for each of these subsets. For brevity, we refer to these as "average home price" and "average rent," respectively. This comparison is shown in Figure 10 for both home price and rent, where we see that *the average home price and rent are higher in ZIP codes with a university than those* *without, and highest in ZIP codes with a medium university*. The pairwise significance of the most recent values (May 2017), calculated using a one-tailed heteroscedastic Student’s *t*-test, is shown in Table 11 for home prices and Table 12 for rent.

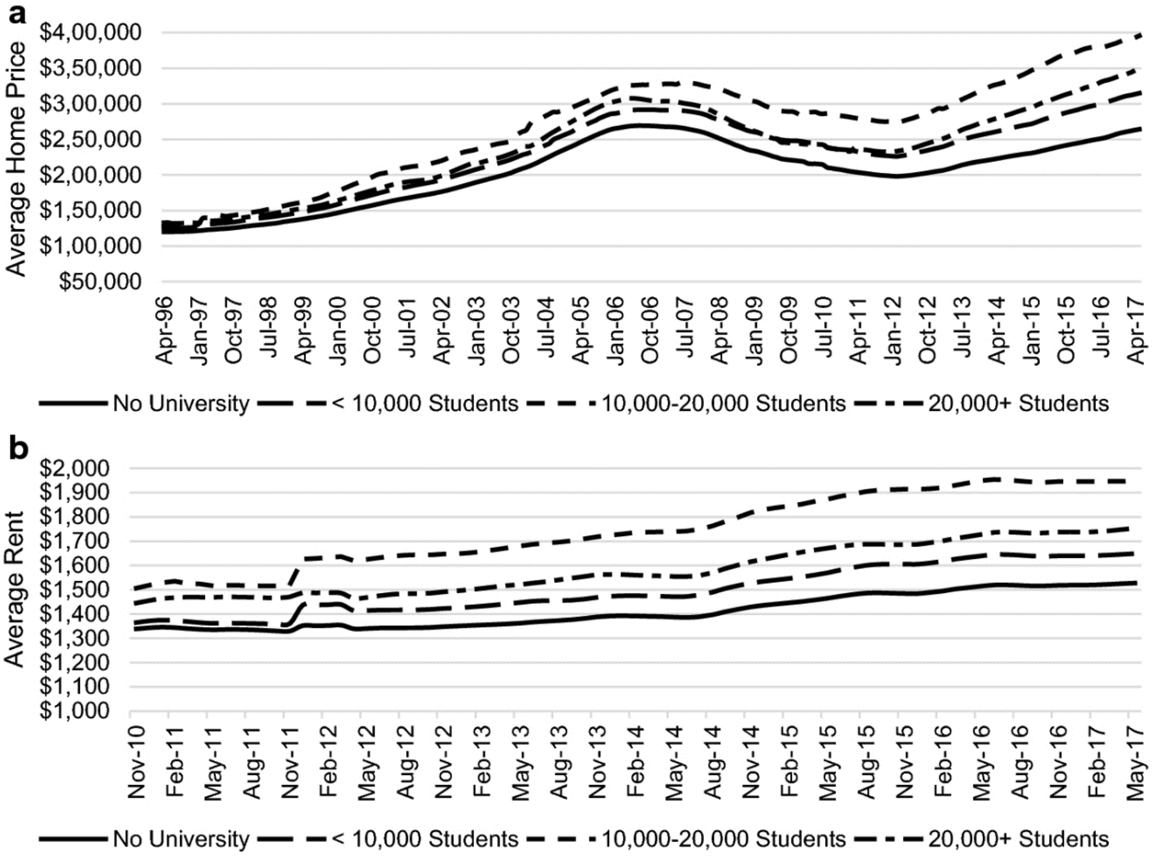


Figure 10. Average home price (a) and rent (b) based on university size.

Table 11. Pairwise significance (*p*-values) of average May 2017 home prices in ZIP codes by university size.

	No university	Fewer than 10,000 students	10,000–20,000 students
Fewer than 10,000 students	0.00361	–	–
10,000-–20,000 students	0.00676	0.0733	–
20,000 or more students	0.0132	0.203	0.239

Table 12. Pairwise significance (*p*-values) of average May 2017 rent in ZIP codes by university size.

	No university	Fewer than 10,000 students	10,000–20,000 students
Fewer than 10,000 students	0.00615	–	–
10,000–20,000 students	0.00118	0.0189	–
20,000 or more students	0.00258	0.128	0.106

Similarly, we compared the average home price and rent over time for four ZIP code subsets based on hospitals. This comparison was between ZIP codes with no hospital, ZIP codes with a small hospital (fewer than 100 affiliated doctors), ZIP codes with a medium hospital (at least 100 but fewer than 500 affiliated doctors) and ZIP codes with a large hospital (500 or more affiliated doctors). This comparison is shown in Figure 11 for both home price and rent, where we see that *ZIP codes with larger hospitals have higher average home price and rent than those with smaller hospitals, while only ZIP codes with large hospitals have higher average home price and rent than ZIP codes with no hospital*. Figure 12 shows the correlations between the number of doctors affiliated with a hospital and both home price (Pearson correlation 0.154) and rent (Pearson correlation 0.261). The p-value for both correlations is less than 1×10^{-5}. The pairwise significance of the most recent home price and rent values (May 2017), calculated using a one-tailed heteroscedastic Student's t-test, is shown in Table 13 for home prices and Table 14 for rent.

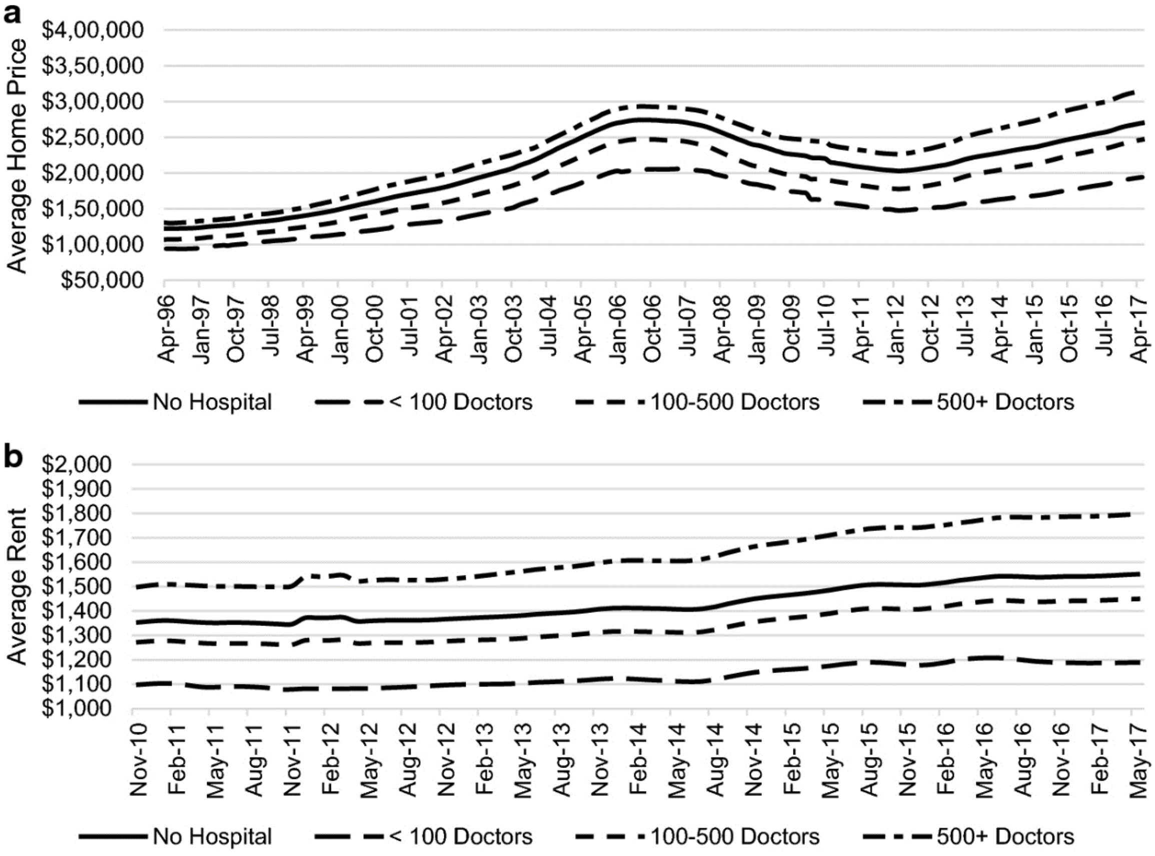


Figure 11. Average home price (a) and rent (b) based on hospital size.

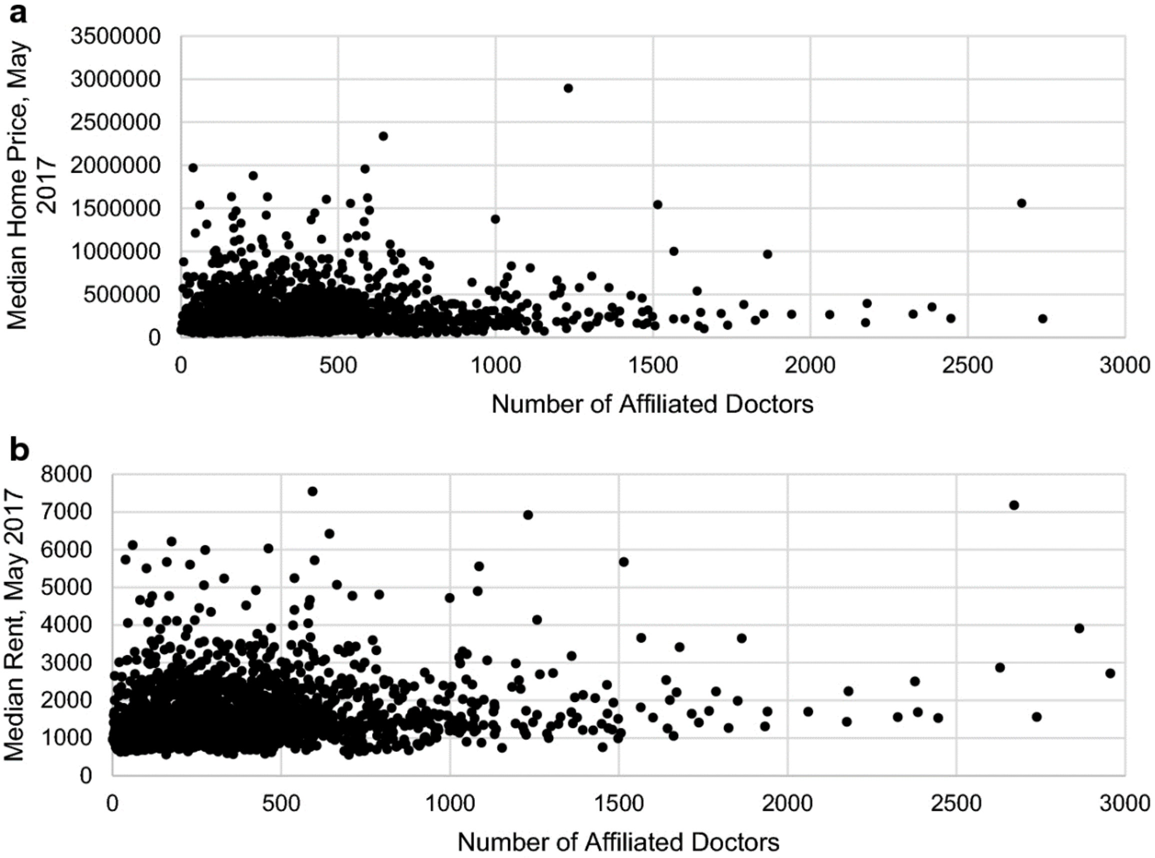


Figure 12. Number of doctors vs. home price (a) or rent (b) in ZIP codes with a hospital.

Table 13. Pairwise significance (*p*-values) of average May 2017 home prices in ZIP codes by hospital size.

	No hospital	Fewer than 100 doctors	100–500 doctors
Fewer than 100 doctors	4.02 × 10−10	–	–
100–500 doctors	7.27 × 10−5	2.85 × 10−5	–
500 or more doctors	0.000637	1.94 × 10−11	2.35 × 10−6

Table 14. Pairwise significance (*p*-values) of average May 2017 rent in ZIP codes by hospital size.

	No hospital	Fewer than 100 doctors	100–500 doctors
Fewer than 100 doctors	1.71 × 10−43	–	–
100–500 doctors	4.11 × 10−9	4.84 × 10−20	–
500 or more doctors	2.72 × 10−9	2.63 × 10−35	4.79 × 10−15

Appreciation of home price and rent

We found several very weak correlations between the number of students enrolled in a university and average annual home price and rent appreciation in ZIP codes with a university. These correlations are listed in Table 15 for home price appreciation and

Table 16 for rent appreciation. For hospitals, we found *a weak positive correlation between the number of doctors affiliated with a hospital and average annual home price appreciation in ZIP codes with a hospital and a population below the national ZIP code average* (Figure 13; Pearson correlation 0.203, *p*-value 0.0016). We also found a very weak correlation between the number of doctors affiliated with a hospital and average annual home price appreciation in all ZIP codes with a hospital (Pearson correlation 0.107, *p*-value $< 1 \times 10^{-5}$) in addition to several very weak correlations between the number of doctors affiliated with a hospital and average annual rent appreciation in ZIP codes with a hospital. These correlations are listed in Table 17.

Table 15. Correlations between the number of students enrolled in a university and home price appreciation.

Subset	**Pearson correlation**	***p*-value**
ZIP codes with a university	0.132	0.000864
ZIP codes with a university and home prices below the national ZIP code average	0.136	0.0063
ZIP codes with a university and home prices above the national ZIP code average	0.115	0.0822
ZIP codes with a university and a population above the national ZIP code average	0.134	0.00126
ZIP codes with a university and population density above the national ZIP code average	0.107	0.0359

Table 16. Correlations between the number of students enrolled in a university and rent appreciation.

Subset	**Pearson correlation**	***p*-value**
ZIP codes with a university	0.124	0.000575
ZIP codes with a university and rent below the national ZIP code average	0.142	0.00232
ZIP codes with a university and a population above the national ZIP code average	0.117	0.00218
ZIP codes with a university and population density below the national ZIP code average	0.144	0.00793

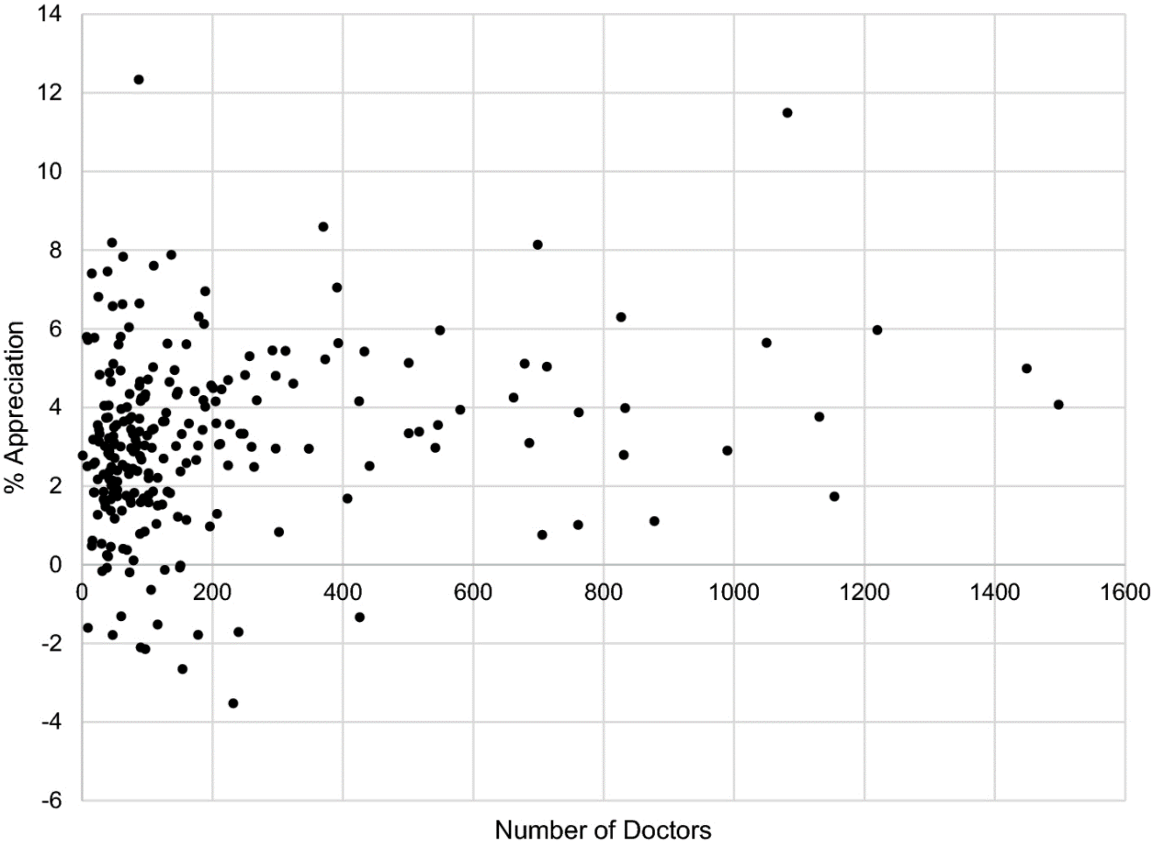


Figure 13. Number of doctors vs. appreciation in hospital ZIPs with population below national ZIP code average.

Table 17. Correlations between the number of doctors affiliated with a hospital and rent appreciation.

Subset	**Pearson correlation**	***p*-value**
ZIP codes with a hospital	0.197	$< 1 \times 10^{-5}$
ZIP codes with a hospital and rent below the national ZIP code average	0.173	$< 1 \times 10^{-5}$
ZIP codes with a hospital with fewer than 1000 affiliated doctors	0.17	$< 1 \times 10^{-5}$
ZIP codes with a hospital and a population below the national ZIP code average	0.156	0.000832
ZIP codes with a hospital and a population above the national ZIP code average	0.167	$< 1 \times 10^{-5}$
ZIP codes with a hospital in a metropolitan area	0.166	$< 1 \times 10^{-5}$

Volatility of Home Price and Rent

We found a weak positive correlation between the number of students enrolled in a university and home price volatility in ZIP codes with a university and a population below the national ZIP code average (Figure 14a; Pearson correlation 0.296, *p*-value

0.0299) as well as several very weak correlations between the number of students enrolled in a university and home price volatility in ZIP codes with a university. These correlations are listed in Table 18. For hospitals, we found a weak positive correlation between the number of doctors affiliated with a hospital and home price volatility in ZIP codes with a hospital and a population below the national ZIP code average (Figure 14b; Pearson correlation 0.244, *p*-value 0.000134). We also found several very weak correlations between the number of doctors affiliated with a hospital and home price and rent volatility in ZIP codes with a hospital. These correlations are listed in Table 19 for home price volatility and Table 20 for rent volatility.

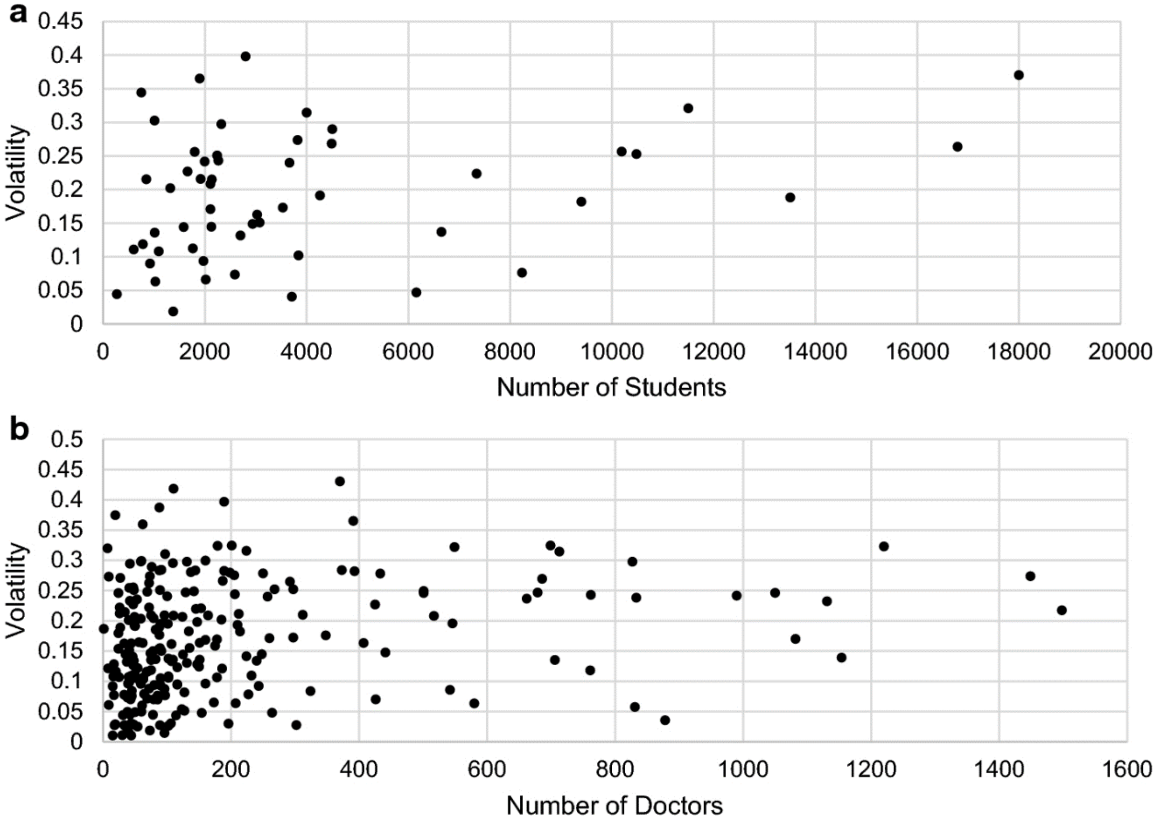


Figure 14. Home price volatility in ZIP codes with population below the national ZIP code average. (a) Number of students vs. home price volatility in ZIP codes with a university. (b) Number of doctors vs. home price volatility in ZIP codes with a hospital.

Table 18. Correlations between the number of students enrolled in a university and home price volatility.

Subset	Pearson correlation	*p*-value
ZIP codes with a university	0.139	0.000449
ZIP codes with a university and home prices below the national ZIP code average	0.1749	0.000434
ZIP codes with a university and a population above the national ZIP code average	0.129	0.00193
ZIP codes with a university and population density below the national ZIP code average	0.124	0.0518

Table 19. Correlations between the number of doctors affiliated with a hospital and home price volatility.

Subset	Pearson Correlation	*p*-value
ZIP codes with a hospital	0.135	$< 1 \times 10^{-5}$
ZIP codes with a hospital with fewer than 1000 affiliated doctors	0.144	$< 1 \times 10^{-5}$
ZIP codes with a hospital and home prices below the national ZIP code average	-0.152	$< 1 \times 10^{-5}$
ZIP codes with a hospital and home prices above the national ZIP code average	-0.108	0.0052
ZIP codes with a hospital and a population above the national ZIP code average	0.103	$< 1 \times 10^{-5}$
ZIP codes with a hospital and population density below the national ZIP code average	0.113	7.6×10^{-5}
ZIP codes with a hospital and population density above the national ZIP code average	-0.106	0.000537

Table 20. Correlations between the number of doctors affiliated with a hospital and rent volatility.

Subset	Pearson correlation	*p*-value
ZIP codes with a hospital	0.105	$< 1 \times 10^{-5}$
ZIP codes with a hospital and a population below the national ZIP code average	0.147	0.00165

Vacancies

We again grouped ZIP codes into four subsets for both universities and hospitals to compare the average percentage of vacant homes between subsets. These comparisons are shown in Figure 15. Among ZIP codes with a university, we see that *the average percentage of vacant homes is highest in ZIP codes with medium universities and lowest in ZIP codes with no university, while ZIP codes with small universities have a higher average percentage of vacant homes than ZIP codes with large universities*. Among ZIP codes with a hospital, we see that *the average percentage of vacant homes is highest in ZIP codes with small hospitals and lowest in ZIP codes with no hospital, while ZIP codes* *with large hospitals have a higher average percentage of vacant homes than ZIP codes with medium hospitals*. The pairwise significance of the most recent values (Q2 2017), calculated using a one-tailed heteroscedastic Student’s *t*-test, is shown in Table 21 for ZIP codes grouped by university size and Table 22 for ZIP codes grouped by hospital size.

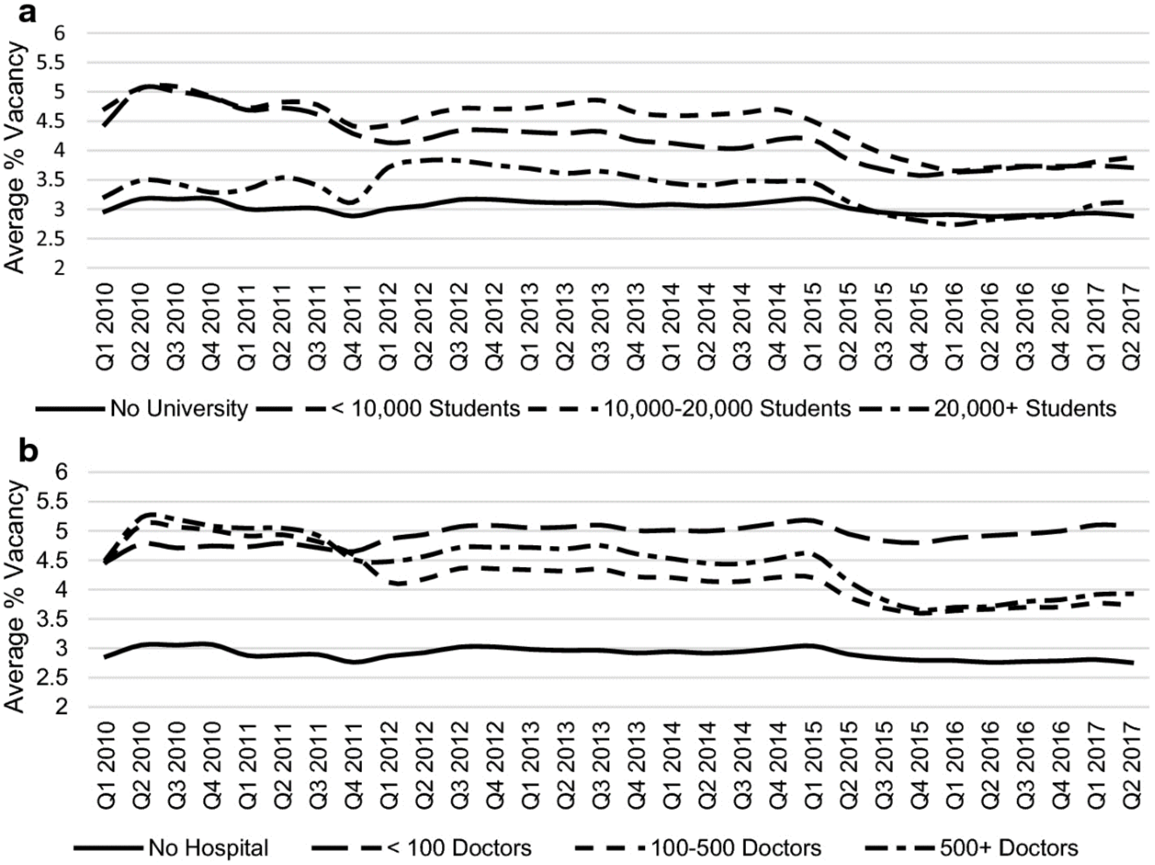


Figure 15. Average percentage of vacant homes based on university (a) and hospital (b) size.

Table 21. Pairwise significance (*p*-values) of average home vacancy percentage in ZIP codes by university size.

	No university	Fewer than 10,000 students	10,000–20,000 students
Fewer than 10,000 students	1.89×10^{-11}	–	–
10,000–20,000 students	0.0013	0.43	–
20,000 or more students	0.362	0.00802	0.0246

Table 22. Pairwise significance (*p*-values) of average home vacancy percentage in ZIP codes by hospital size.

	No hospital	Fewer than 100 doctors	100–500 doctors

Fewer than 100 doctors	1.3×10^{-66}	–	–
100–500 doctors	8.55×10^{-48}	1.13×10^{-21}	–
500 or more doctors	4.9×10^{-15}	3.07×10^{-11}	0.206

We found a weak positive correlation between the number of students enrolled in a university and the percentage of vacant homes in ZIP codes with a university and a population below the national ZIP code average (Figure 16; Pearson correlation 0.285, *p*-value 0.0368). However, we also found a very weak negative correlation between the number of students enrolled in a university and the percentage of vacant homes in ZIP codes with population density below the national ZIP code average (Pearson correlation − 0.134, *p*-value 0.0361). Among ZIP codes with a hospital, we found very weak correlations between the number of doctors affiliated with a hospital and the percentage of vacant homes in ZIP codes with home prices above the national ZIP code average (Pearson correlation 0.14, *p*-value 0.000296) and rent above the national ZIP code average (Pearson correlation 0.129, *p*-value 0.000134).

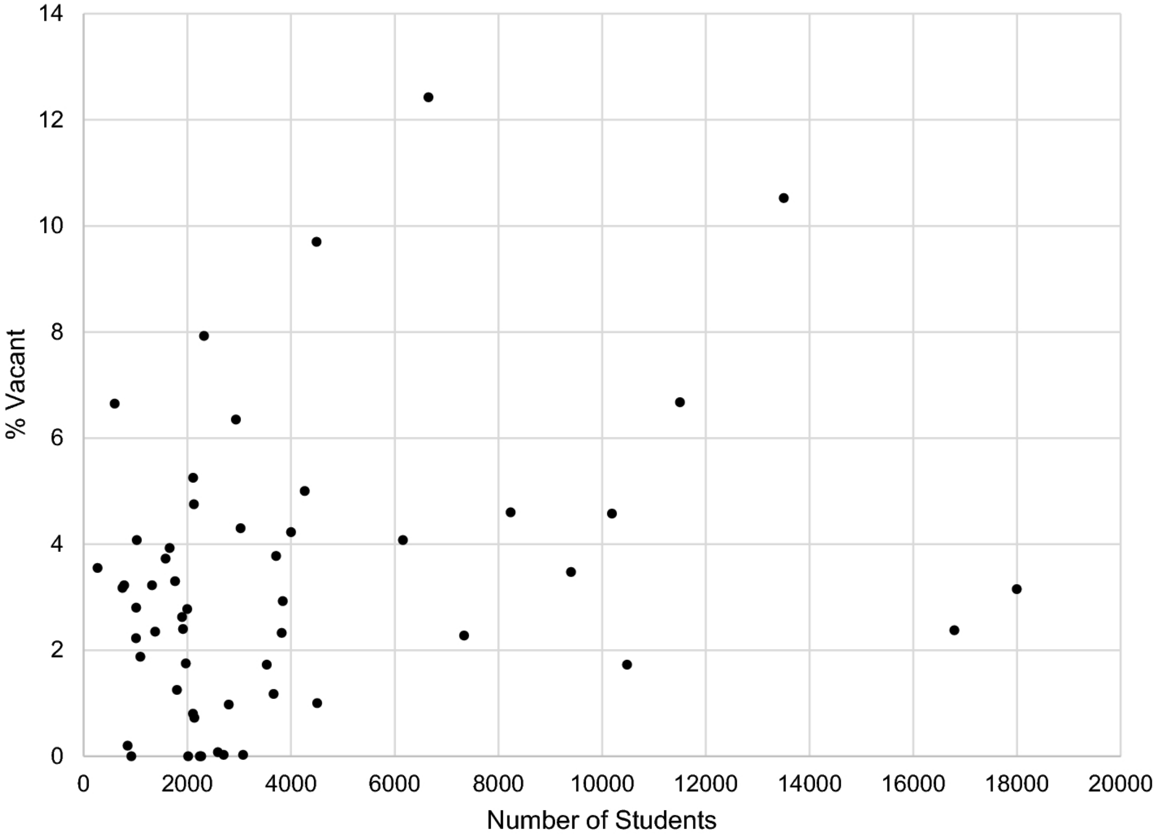


Figure 16. Percentage of vacant homes in university ZIP codes with population below national ZIP code average.

University Rankings

We found *weak negative correlations between university ranking and both home price and rent.* The Pearson correlation is −0.269 for university ranking vs. home price with a *p*-value of 0.021. The Pearson correlation is −0.327 for university ranking vs. rent with a *p*-value of 0.00271. These results are not surprising as the negative correlations imply that real estate prices tend to be higher in ZIP codes with higher ranked universities.

3.4.2 Home-Level Analysis

Our goal is to determine whether there exist subsets of the data, partitioned across the dimensions of the university, hospital and home data, in which the distance to a university or hospital is significantly correlated to home price or rent. A few possible avenues for finding these subsets are along the features of the data such as distance ranges, number of bedrooms, types of university and number of doctors affiliated with hospitals. We try to filter data layer by layer by using the feature filters to arrive at a particular high-correlation subset of data.

As discussed in Methods section, each entry in the data table consists of details for homes within ten miles of a university along with that university's details. If a home is near multiple universities, only the entry with the shortest distance from a university is considered. We applied the same scheme to home-hospital data. Table 23 shows the average number of homes for sale and for rent within ten miles of a university or hospital.

Table 23. Average number of homes near a university or hospital (within ten-mile radius).

Home type	**University**	**Hospital**
For rent	165.028	497.483
For sale	107.686	315.817

Analysis of University and Hospital Proximity on Home Price and Rent

As a preliminary analysis, we analyzed the effective distance range to which the presence of a university affects home prices and rent. Table 24 shows the correlations between home price/rent and distance from the nearest university based on different maximum distances. Although all such correlations are very weak, we observed slightly higher correlations for both home price and rent among homes within two miles of a university. Therefore, unless mentioned otherwise, all further experiments related to homes near universities limit the dataset to homes that are within two miles of the nearest university.

Table 24. Distribution of homes by distance from nearest university and home price/rent-distance correlation.

Miles from a university	Homes with home price data	Pearson correlation (home price-distance)	Homes with rent data	Pearson correlation (rent-distance)
< 2	129,102	−0.149	115,340	−0.087
< 3	228,185	−0.122	153,668	−0.076
< 4	326,055	−0.101	183,404	−0.055
< 5	414,769	−0.095	205,884	−0.038
> 5 and < 10	290,713	−0.017	61,602	0.016

Next, we examined at the effect of proximity to a hospital on home prices and rent. As a preliminary experiment, we calculated the home price-distance and rent-distance correlations based on different maximum distances and found the best correlations by partitioning the home data at three miles for home prices and two miles for rent. As seen in Table 25*, homes within a three-mile radius of a hospital have higher correlation between home price and distance from a hospital*. In the remainder of this section, we consequently focus on other data filters based on the number of bedrooms in a home and the number of doctors affiliated with a hospital to find correlations between home price and distance from a hospital. Table 25 also shows that for rent data, the highest correlation between rent and distance from the nearest hospital exists beyond a two-mile radius from the hospitals. Interestingly, the correlation is positive, that is, the rent is higher for homes farther from a hospital.

Table 25. Home price/rent-distance from hospital correlations for various distance ranges up to ten miles.

Home price-distance analysis			Rent-distance analysis		
Miles from a hospital	Number of homes	Pearson correlation	Miles from a hospital	Number of homes	Pearson correlation
Any	877,067	−0.0495	Any	300,768	0.079
< 3 miles	303,313	−0.081	< 2 miles	123,313	−0.013
> 3 miles	573,754	0.006	> 2 miles	177,455	0.094

Analysis of University/Hospital Proximity: Price and Rent Analysis by Number of Bedrooms

For these experiments, we partitioned the home data based on the number of bedrooms. In the first of these experiments, we analyzed the correlations between home price/rent and distance from a university within two miles of a university. We found that *two-bedroom homes have the highest correlation between home price and distance from a university* (Pearson correlation −0.319). This was followed closely by one-bedroom homes. The correlation was very weak for homes with more than two bedrooms. We also found a *weak correlation between rent and distance from a university for one-bedroom homes* (Pearson correlation −0.191). The Pearson correlations for various numbers of bedrooms and the home counts for each such category are shown in Table 26.

Table 26. Home price/rent-distance from university correlations based on distribution by number of bedrooms.

Number of bedrooms	**Homes with home price data (< 2 miles)**	**Pearson correlation (home price-distance)**	**homes with rent data (< 2 miles)**	**Pearson correlation (rent-distance)**
1	13,663	−0.219	32,473	−0.191
2	31,050	−0.319	40,371	−0.142
3	44,028	−0.225	26,369	−0.149
4	24,206	−0.212	11,566	−0.056
> 4	16,155	−0.119	4563	0.073

Next, we analyzed the correlations between home price/rent and distance from a hospital. As shown in our earlier experiments, the set of homes less than three miles away from the nearest hospital is a good candidate for analyzing the effect of proximity to a hospital on home prices, while the set of homes more than two miles away from the nearest hospital is a good candidate for analyzing the effect of proximity to a hospital on rent. For home price data, Table 27 shows that single-bedroom homes have a higher correlation between home price and distance from a hospital (Pearson correlation −0.223) than the other bedroom categories. In the next subsection, we shall thus focus only on these one-bedroom homes. Among the correlations between rent and distance from the nearest hospital for each category, the correlations get stronger as the number of bedrooms increases, as shown in Table 27. The strongest of these is a weak positive correlation for homes with more than four bedrooms (Pearson correlation 0.186).

Table 27. Home price/rent-distance from hospital correlations based on distribution by number of bedrooms.

Number of bedrooms	**Homes with home price data (< 3 miles)**	**Pearson correlation (home price-distance)**	**Homes with rent data (> 2 miles)**	**Pearson correlation (rent-distance)**
1	20,524	−0.223	24,102	−0.008
2	61,831	−0.140	55,195	0.015
3	114,657	−0.100	63,904	0.055
4	68,914	−0.095	26,151	0.114
> 4	37,837	−0.048	2935	0.186

Analysis of University Proximity: Price and Rent Analysis by Type and Rank of University

For home price analysis within two miles of a university, we classify the universities into the following three types: public, private and other. Also, as observed in previously, two-bedroom homes near universities provide a good enough correlation to be explored further. Table 28, which compares the correlations between home price and distance for these types of universities, shows that *two-bedroom homes have a weak negative correlation between home price and distance from a private university within a two-mile radius.*

Table 28. Home price/rent-distance correlations based on university type distribution.

University type	**University count**	**Homes with home price data (< 2 miles, 2 bedrooms)**	**Pearson correlation (home price-distance)**	**Homes with rent data (< 2 miles, 1 bedroom)**	**Pearson correlation (rent-distance)**
Private	1068	16,680	−0.368	16,992	−0.311
Public	548	7812	−0.220	6463	−0.203

For rent analysis within two miles of a university, we again classify the universities into the three types mentioned previously and limit our experiment to one-bedroom homes due to the higher correlations found with those homes. In Table 28, we found a *negative correlation between rent and the distance from a university for one-bedroom homes near a private university* (Figure 17; Pearson correlation −0.311). We also observed a weaker correlation for one-bedroom homes near public universities. Note that we have omitted results for homes near "other" universities since there were very few of these universities compared to the other two types and they yielded very small correlations.

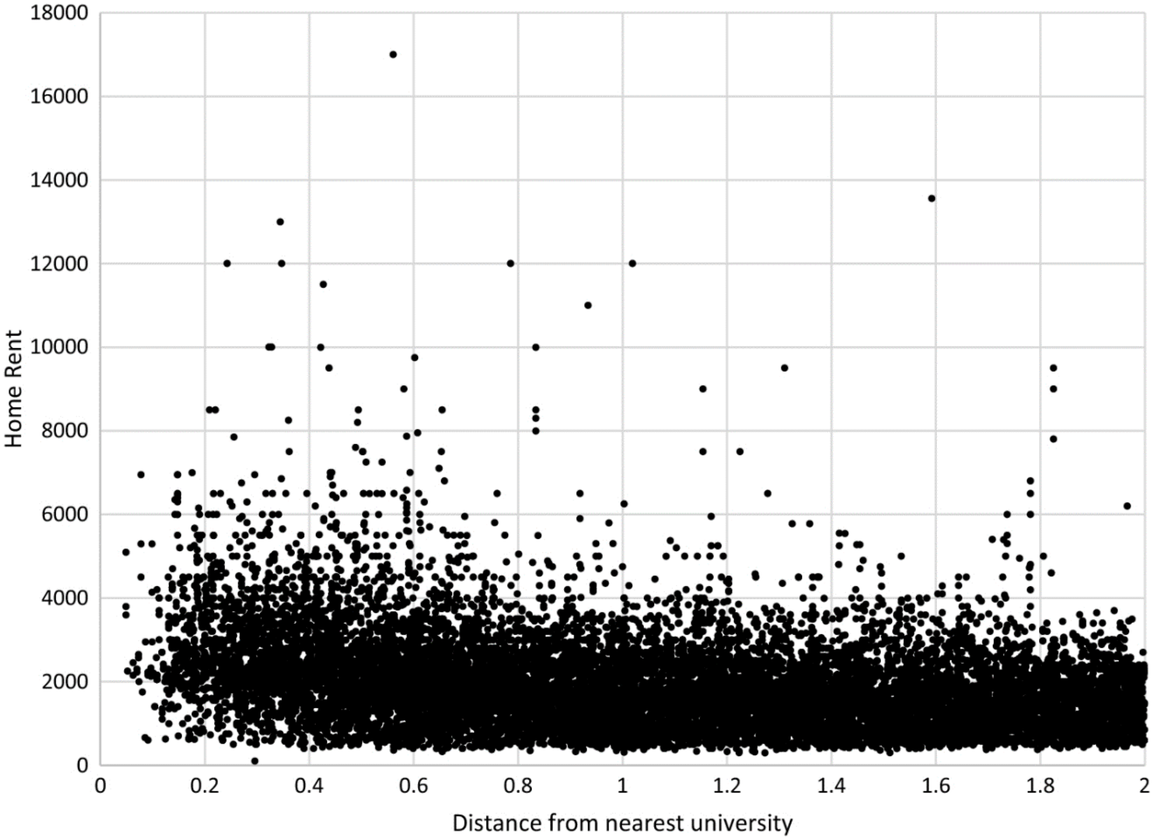


Figure 17. Distance from nearest university vs. rent for homes near a private university.

We also considered the rank of universities in our analysis. The university rankings provided by US News and World Report provide data for only the top 200 schools. However, we found no significant correlations in our experiments involving university rankings as can be seen in Table 29.

Table 29. Correlation of distance from university to home price/rent for top ranked/unranked universities.

	Distance-rent correlation	**Distance-price correlation**
Top 200 ranked universities	−0.007	−0.055
Unranked universities	−0.005	−0.039

We then checked for any interesting correlations between home price/rent and distance from university on the filters of ranked or unranked universities for homes within a two-mile radius. Further, on filtering over one-bedroom homes for rent and two-bedrooms for home price, we found similar correlation for ranked as well as unranked universities. Results for these experiments are shown in Table 30 and Table 31. Hence, as per our analysis, the ranking of a university does not play a crucial role in the dynamics of real estate prices of nearby homes.

Table 30. Correlation of university distance (within two miles) to home price/rent for top ranked/unranked universities.

	Distance-rent correlation	**Distance-price correlation**
Top 200 ranked universities	−0.052	−0.137
Unranked universities	−0.105	−0.160

Table 31. Correlation of university distance (within two miles) to two-bedroom home price/one-bedroom rent for top ranked/unranked universities.

	Distance-rent correlation (one-bedroom homes)	**Distance-price correlation (two-bedroom homes)**
Top 200 ranked universities	−0.230	−0.298
Unranked universities	−0.180	−0.327

Analysis of Hospital Proximity: Price and Rent Analysis by Number of Affiliated Doctors

For home price data, we consider only single-bedroom homes as they exhibited the highest correlation between home price and distance from the nearest hospital. Table 32 shows that *single-bedroom homes near larger hospitals (more than 500 doctors) have a higher distance-home price correlation compared to those near smaller hospitals.*

Table 32. Home price/rent-distance correlations by number of doctors affiliated with the nearest hospital.

Number of doctors	**Homes with home price data (1 bedroom, < 3 miles)**	**Pearson correlation (home price-distance)**	**Homes with rent data (> 4 bedrooms, > 2 miles)**	**Pearson correlation (rent-distance)**
< 500	12,976	−0.196	5168	0.214
> 500	7548	−0.293	2935	−0.085

For rent data, we restricted our analysis to homes with more than four bedrooms at distance of over two miles from the nearest hospital. We then categorize this data into two subsets of fewer than 500 or more than 500 doctors affiliated with the nearest hospital. Table 32 shows that *larger homes (more than four bedrooms) near a smaller hospital (fewer than 500 doctors) had a significantly higher rent-distance correlation as compared to homes near a larger hospital.*

3.5 Discussion

In our analysis of average ZIP code median home price and median rent over time (“average home price” and “average rent”), we found that the average home price and rent are higher in ZIP codes with a university than those without, and highest in ZIP codes with a medium-sized university (10,000–20,000 students). One possible explanation for this observation is that public universities tend to have a more positive effect on home price and rent, as most medium-sized universities in our analysis are public. We also found that ZIP codes with larger hospitals have higher average home price and rent than those with smaller hospitals, while only ZIP codes with large hospitals have higher average home price and rent than ZIP codes with no hospital. One possible reason why ZIP codes with small and medium hospitals have lower home price and rent than ZIP codes with no hospital is that smaller hospitals tend to be in more remote areas with lower real estate prices. In general, these measures were positively affected by the presence of a university and negatively affected by the presence of a hospital (this should not be confused by the impact of the hospital distance of individual homes within a ZIP code). Note that the existence of a large (or small) university in a ZIP code does not imply the existence of a large (or small) hospital or vice versa (Table 33).

Table 33. Number of ZIP codes with both a university and a hospital for each subset.

	No university	**Fewer than 10,000 students**	**10,000–20,000 students**	**20,000 or more students**
No hospital	N/A	305	43	32
Fewer than 100 doctors	290	16	2	1
100–500 doctors	1340	122	24	10
500 or more doctors	404	58	4	15

The strongest ZIP code-level correlations discovered in this study were found for smaller ZIP codes (population below the national ZIP code average). The reason may be that institutions have a higher impact in smaller ZIP codes as they are one of the main employers or drivers of economic activity. Specifically, we found that *in smaller ZIP codes with at least one hospital, there is a positive correlation (0.203) between the number of affiliated doctors and home price appreciation.* This result, along with several weaker correlations we found between home price/rent and appreciation, agrees with our expectation that appreciation is higher near larger institutions. Our analysis of volatility in smaller ZIP codes showed that for ZIP codes with at least one university, there is a positive correlation (0.296) between the number of enrolled students and home price volatility, and for ZIP codes with at least one hospital there is a positive correlation (0.244) between the number of affiliated doctors and home price volatility. *These results on volatility are opposite from what we expected, as larger universities or hospitals* *generally imply more job security for the area, and hence one would expect lower price volatility as well.* We also found that smaller ZIP codes with at least one university have a positive correlation (0.285) between the number of students enrolled and the percentage of vacant homes. This agrees with our expectation that the vacancy rate is higher near larger universities, as many students leave for the summer.

Our analysis of homes near universities or hospitals based on the number of bedrooms in homes showed several interesting correlations. We found that the correlation between home price and distance from a university is strongest for two-bedroom homes (−0.319), while the correlation between rent and distance from a university is strongest with one-bedroom homes (−0.191). *That is, smaller homes are of higher demand closer to universities.* This conclusion seems logical as most of the occupants within a two-mile radius from a university would be students and not big families.

Similarly, we found that the correlation between home price and distance from a hospital is strongest for one-bedroom homes (−0.223), which could imply high demand for single bedroom homes near hospitals. In contrast, the correlation between rent and distance from a hospital was strongest for homes with more than four bedrooms (0.186), which implies that larger families may prefer to live farther from a hospital.

We found negative correlations between the price of a two-bedroom home and distance from a private university (−0.368) or a public university (−0.220). We also found negative correlations between the rent for a one-bedroom home and the distance from a private university (−0.311) or a public university (−0.203). *A probable cause for the difference in correlations between public and private universities is that private* *university students may be willing to pay more rent to be closer to the university.* These results also show that renting a home near a university has a slightly lower correlation compared to the sale of a home, implying a higher demand for buying a home. This may be accounted for by sales to investors for the purposes of renting out these homes. This possibility may be a subject for future research.

As noted in previous sections, economic laws as viewed from the lens of homes lying in the proximity of universities and hospitals act in subtle ways. What seems to be true near a university may not be true near a hospital. Indeed, one should not be altogether surprised by those findings. Although both universities and hospitals are magnets of highly educated workforce, universities have students while hospitals generally do not (with the exception of teaching hospitals, which are by definition universities). Demand for housing is a function of multiple factors which are not altogether easy to decouple—variations in demand differ according to factors that would appeal to different demographic and economic strata. For example, students fuel demand for inexpensive housing lying in close proximity to a university campus. On the other hand, hospitals' professional staff, some highly paid (doctors, senior nurses and senior management), are adult, mostly with families that compete for larger homes, in neighborhoods having amenities commensurate to their needs and desires. Clearly the differences between those two demographical strata are stark. That said, there are many examples of universities that are situated in what may be considered as "inner city" and those include some of the finest universities in this country, e.g. University of Pennsylvania and Temple University (both in Philadelphia), University of Southern California (Los Angeles), Wayne State

University (Detroit), etc., where this analysis would prove wrong. More often, there are many examples of hospitals situated in what one would consider a “bad” part of town, where the professional staff does not live; where doctors, nurses, management, etc. drive to work, sometime for an hour one-way, “put their time” and drive back to their home in a middle- or upper-middle-class suburbia. It is also interesting to note that “job security” plays a secondary role, if that. Indeed, those “old” notions of job security do not seem to play prominently into the economic calculus, especially as it manifests in real estate terms. However, as expected, what is confirmed by the analysis is the notion that demand for modest rent housing is high near an employer promising job security.

3.5.1 Limitations

When considering the distance between homes and universities/hospitals, we used the geographical distance without regard to elevation or roads. The Google Maps API could be used to account for these, but the API rate limits imposed by Google made this impractical. The university rankings provided by US News and World Report provide data for only the top 200 schools. For that, we generally study them in two groups, ranked and unranked.

The CMS hospital data includes smaller medical centers in addition to traditional hospitals. These medical centers tend to have very few affiliated doctors, which may affect our calculations involving subsets of ZIP codes that contain these medical centers. However, these medical centers are often in small cities with no other hospital nearby, thus we believe they are appropriate for our analysis.

Two limitations apply to the Zillow data. First, the prices/rent are based on listed prices/rent and not actual sale prices/rent. Second, the median monthly home price and rent data provided by Zillow had 1 or more months of data missing for some ZIP codes. To account for a ZIP code has one or more consecutive months of missing data between months with data, we assume the change in home price or rent is linear during the months with missing data. If a ZIP code's first month of data is after the first month of Zillow data (April 1996 for home prices and November 2010 for rent), that ZIP code is not included in our calculation of average median home price/rent for months before that ZIP code's first month of data, and our calculation of appreciation and volatility for that ZIP code are made using only the range of months for which we have data for that ZIP code.

We assume a ZIP code containing a university or a hospital contained that institution throughout the entire range of dates used in calculations for that ZIP code; however, some universities or hospitals may have been built after the start of their containing ZIP codes' ranges of home price/rent data.

As discussed above, many factors affect the demand–and therefore the price–of housing. While our study focuses on a select few factors, our home price and rent data may be affected by one or more other variables that we do not consider.

3.6 Conclusions

We analyzed several measures of real estate valuation near universities and hospitals based on both individual home sales and ZIP code level aggregates. In our ZIP code-level analysis, we found that ZIP codes with universities tend to have above average median home price and median rent, especially those with medium-sized universities, while ZIP codes with hospitals tend to have below average median home price and median rent, with the exception of those with large hospitals, and that less populated ZIP codes have positive correlations between the number of doctors affiliated with a hospital and home price appreciation, and between the number of enrolled university students and home vacancy rate. Notably, less populated ZIP codes also have positive correlations between home price volatility and both the number of enrolled students (in ZIP codes with a university) and the number of affiliated doctors (in ZIP codes with a hospital), which is surprising given that one would expect these institutions to have a stabilizing effect on home prices. In our home-level analysis, we found that the home price and rent for smaller homes tend to be the most affected by distance from a university, while distance from a hospital has greater effect on both the price of one-bedroom homes as well as on the rent of large homes. Of particular interest is our finding of a positive correlation between rent and distance from a hospital beyond two miles, suggesting that renters prefer homes in areas without a hospital nearby.

The findings point at complex interactions between demand and supply in the ZIP codes and homes under study. There is little doubt that supply–demand curves should be stratified by price points and possibly additional factors. This is clearly demonstrated in the city of Irvine, California, (ZIP code 92618) where two large healthcare facilities, Kaiser and Hoag hospitals, employ a large staff at a diverse income levels: from board-certified surgeons at the higher end, to nurse assistants and orderlies at the other. As one may readily check on Zillow or similar websites, there is little, if any "affordable" housing in the vicinity of ZIP code 92618, presumably necessitating low-income hospital staff to seek housing in lower-rent areas. An overall theory to explain behavior of real estate in the vicinity of a university or a hospital may prove complex as it should take into account myriad hard-to-measure factors. We will take this kind of analysis in a subsequent study, specifically the effects of interactions between economics, demographics, and amenities, to further explore how all the effects interact with the metrics we normally associate with real estate and potentially develop a machine learning model based on these analyses.

Chapter 4

Classification of Health-Related Social Media Posts: Evaluation of Post Content Classifier Models and Analysis of User Demographics

Background: The increasing volume of health-related social media activity, where users connect, collaborate, and engage, has increased the significance of analyzing how people use health-related social media.

Objective: The aim of this study was to classify the content (e.g. posts that share experiences and seek support) of users who write health-related social media posts and study the effect of user demographics on post content.

Methods: We analyzed two different types of health-related social media: (1) health-related online forums—WebMD and DailyStrength—and (2) general online social networks—Twitter and Google+. We identified several categories of post content and built classifiers to automatically detect these categories. These classifiers were used to study the distribution of categories for various demographic groups.

Results: We achieved an accuracy of at least 84% and a balanced accuracy of at least 0.81 for half of the post content categories in our experiments. In addition, 70.04% (4741/6769) of posts by male WebMD users asked for advice, and male users' WebMD posts were more likely to ask for medical advice than female users' posts. The majority of posts on DailyStrength shared experiences, regardless of the gender, age group, or

location of their authors. Furthermore, health-related posts on Twitter and Google+ were used to share experiences less frequently than posts on WebMD and DailyStrength. Conclusions: We studied and analyzed the content of health-related social media posts. Our results can guide health advocates and researchers to better target patient populations based on the application type. Given a research question or an outreach goal, our results can be used to choose the best online forums to answer the question or disseminate a message.

4.1 Introduction

4.1.1 Background

There is a huge amount of knowledge waiting to be extracted in health-related online social networks and forums, which we collectively refer to as social media. Health-related social media store the interactions of users who are interested in health-related topics [52]. These users share their experiences, share information of friends and family, or seek help for a wide range of health issues [52]. In the United States, more than 60 million Americans have read or collaborated in *health 2.0* resources [53]. In addition, 40% of Americans have doubted a professional opinion when it conflicted with the opinions expressed in health-related social media [53]. Health-related social media widen access to health information for the public, regardless of individuals' race, age, locality, or education [52].

In this study, we evaluated the content of posts in various health-related social media. We analyzed two types of health-related social media: (1) health-related online forums: WebMD and DailyStrength and (2) general social networks: Google+ and Twitter. This was a 4-step process comprising data collection, identifying post content categories, performing classification experiments, and performing a demographics analysis. We first collected large datasets of posts from each source and identified several categories. Afterward, we identified meaningful categories from randomly selected posts from each source. In our classification experiments, we labeled data from each source and trained classifiers to identify post content categories. Finally, we used classifiers trained on our labeled data to identify categories in the remaining data and analyzed how often posts in these categories are made by various demographic groups.

The goal of this study was to provide researchers with information and tools to support further research. For example, researchers looking for clinical trial participants can use DailyStrength, where users often share experiences about a particular condition, and health advocates seeking to spread awareness about a condition that affects men can use WebMD, where men often ask for advice. To this end, we also made comparisons between platforms to suggest where such a researcher might begin looking. The classifier models built in this study can assist with this task as well as other analyses involving health-related online postings.

4.1.2 Related Work

Analysis of Health-Related Social Media

Many studies have been performed to characterize health-related social media communities. Hackworth and Kunz [54] reported that 80% of Americans have searched the internet for health-related information, more than 60 million Americans are consumers of social networks in the Web 2.0 environment (health 2.0), and consumers, especially those with chronic conditions, are leading the health 2.0 movement by seeking clinical knowledge and emotional support. Wiley et al. [55] studied the impact of different characteristics of various social media forums on drug-related content and demonstrated that the characteristics of a social media platform affect several aspects of discussion. Eichstaedt et al. [56] predicted the county-level heart disease mortality by capturing the psychological characteristics of local communities through expressed text in Twitter. However, these studies do not describe or compare specific demographics in terms of their post content.

Further work has focused on categorizing health-related posts based on their content. Yu et al. [57] performed a preliminary content analysis of D/deaf and hard of hearing discussion forum AllDeaf to observe different types of social support behaviors and identify social support features for a future text classification task. Reavley and Pilkington [58] analyzed the content of tweets related to depression and schizophrenia, finding that tweets about depression mostly discussed consumer resources and advertisements, whereas tweets about schizophrenia mostly raised awareness and reported research findings. Lee et al. [59] analyzed the content of tweets from health-related Twitter users, finding that they tweet about testable claims and personal experiences. Lopes and da Silva [60] collected posts from a health-related online forum, MedHelp, and used them to propose and refine a scheme for manually classifying health-related forum posts into 4 categories and a total of 23 subcategories. Our work was built upon these studies by defining our own categories of post content, some of which have analogues in these studies.

Health-Related Demographic Analysis

Other work has compared health issues between demographics or examined the demographics within a population participating in health-related research. Krueger et al. [61] studied the mortality attributable to a low education level in the United States across several demographics, where they found people with an education level below a high school degree to have a higher mortality rate. Anderson-Bill et al. [62] examined the demographics and behavioral and psychosocial characteristics of *Web-health users* (adults who use the Web to find information on health behavior and behavior change) recruited for a Web-based nutrition, physical activity, and weight gain prevention intervention. Their results suggest that users participating in online health interventions are likely “middle-aged, well-educated, upper middle-class women whose detrimental health behaviors put them at risk of obesity, heart disease, some cancers, and diabetes” [62]. These studies describe the demographics of the populations in their studies but do not describe the demographics of health-related social media users.

Previous work has focused on characterizing demographics on health-related social media. Sadah et al. [63] analyzed the demographics of health-related social media and found that users of drug review websites and health-related online forums are predominantly women, health-related social media users are generally older than general social media users, black users are underrepresented in health-related social media, users in areas with better access to health care participate more in health-related social media, and the writing level of health-related social media users is lower than the reading level of the general population. Sadah et al. [64] also performed a demographic-based content analysis of health-related social media posts to extract top distinctive terms, top drugs and disorders, sentiment, and emotion, finding that the most popular topic varied by demographic, for example, pregnancy was popular with female users, whereas cardiac problems, HIV, and back pain were the most discussed topics by male users. They also found that users with a higher writing level were less likely to express anger in their posts. We expanded upon this work by characterizing and comparing the demographics of health-related social media websites in terms of the frequency of post content categories.

Text Classification in Social Media

Text classification is frequently employed by researchers to gain insights into social media users and trends, both in and out of health-related settings. Sadilek et al. [65] studied the spread of infectious diseases by analyzing Twitter data using a support vector machine (SVM) model. Huh et al. [66] developed a naïve Bayes model to help WebMD moderators find posts they would likely respond to. Nikfarjam et al. [67] proposed a machine learning-based tagger to extract adverse drug reactions from health-related social media. Mislove et al. [68] estimated the gender and ethnicity of Twitter users using the reported first name and last name. Sadah et al. [63] expanded upon the work of Mislove et al. [68] by considering screen names in estimating gender. In this study, we used text classification techniques to identify categories of post content in health-related social media and used the techniques proposed in the studies by Sadah et al. [63] and Mislove et al. [68] to study the frequency of these categories within several demographics.

4.2 Methods

4.2.1 Datasets

For health-related online forums, we selected 2 different websites, WebMD and DailyStrength. The reason for selecting 2 health-related online forums is to cover the different types of health-related online forums that they each represent. Although WebMD consists of multiple health communities where people ask questions and get responses from the community members [69], DailyStrength enables patients to exchange experiences and treatments, discuss daily struggles and successes, and receive emotional support [70]. For each post collected from these websites, we extracted the URL, title, author's username, post time, the body of the post, and the name of the message board. For each user of a collected post, we also collected the author's age, friends, gender, and location, where applicable. As crawling of these sites has been performed at different times, some of the data we have collected do not reflect the current availability of certain attributes because of website format changes, for example, age and gender are currently available from WebMD user profiles but were not available before. In this study, the selection of demographic attributes we used for a source is based on the availability reflected by the majority of posts collected from that source, for example, most of the WebMD posts in our data were collected before age and gender were available, thus we did not use these attributes for an analysis of WebMD user demographics. We restricted the posts used from these sources to the first post in each thread. In our analysis, we used the post body, post title, message board name, and username from WebMD and the post body, post title, message board name, and user’s gender, age, and location from DailyStrength.

For general social networks, we chose Twitter and Google+ as they offer interfaces to easily collect their data (in contrast to Facebook). For each Twitter post, we collected the post content, post time, location, and the author’s username and location. For each Google+ post we collected the title, post time, update time, the post content, the location, and the author’s username, first and last names, age, gender, and location. As Twitter and Google+ are general social networks, we used 274 representative health-related keywords to filter them as follows: (1) Drugs: from the most prescriptions dispensed from RxList [71], we selected the 200 most popular drugs. By removing the variants of the same drug (e.g. different milligram dosages), the final list of drugs contained 124 unique drug names. (2) Hashtags: 11 popular health-related Twitter hashtags, such as #BCSM (Breast Cancer and Social Media). (3) Disorders: 81 frequently discussed disorders, such as AIDS and asthma. (4) Pharmaceuticals: the names of the 12 largest pharmaceutical companies, such as Novartis. (5) Insurance: the names of the 44 biggest insurance companies, such as Aetna and Shield. (6) General health-related keywords “healthcare” and “health insurance.” To reach the final keyword counts for hashtags, disorders, pharmaceuticals, and insurance, we sampled each keyword from a larger list for each of these categories and kept keywords with a high ratio of health-related posts. In our analysis, we used the tweet body, user’s first and last name, and user’s location from Twitter and post body, post title, and user’s gender, age, first and last name, and location from Google+.

To filter Twitter with the health-related keyword list to retrieve relevant tweets for TwitterHealth, we used the Twitter streaming application programming interface (API) [72]. Similarly, we used Google+ API [73] to extract the relevant posts for Google+Health. For health-related online forums WebMD and DailyStrength, we built a crawler for each website in Java using jsoup [17], a library to extract and parse HTML content. Table 34 lists for each source the number of posts collected, the date ranges of collected posts, and whether the demographic attributes used in this study are present, and Table 35 lists the distribution of demographics for each source across each demographic attribute. For all 4 of these sources, we did not specifically focus our search on English-language posts aside from using English drug names; however, the majority of posts collected from these sources were in the English language.

Table 34. List of all sources used with their number of posts, date range of posts, and the available demographic attributes.

Source	**Number of posts**	**Date range**	**Gender**	**Age**	**Ethnicity**	**Location**
TwitterHealth [74]	11,637,888	May 2, 2013 to November 11, 2013	Gender classifier [68]	No[7]	Ethnicity classifier [68]	Yes[8]
Google+Health [75]	186,666	August 24, 2009 to January 5, 2014	Yes	Yes	Ethnicity classifier [68]	Yes
DailyStrength [76]	1,319,622	June 21, 2006 to December 3, 2017	Yes	Yes	No	Yes
WebMD [77]	318,297	December 24, 2006 to May 11, 2019	Gender classifier [63]	No	No	No

[7] The demographic attribute is not provided by the source and no classifier is used because of low accuracy.
[8] The demographic attribute is provided by the source.

Table 35. Demographics of users from each source.

Attribute and demographic	**TwitterHealth, %**	**Google+Health, %**	**DailyStrength, *n* (%)**	**WebMD, *n* (%)**
Gender				
Male	48.19[9]	64.64[9]	95,269 (17.26)[10]	6769 (32.41)[10]
Female	51.81[9]	35.36[9]	456,600 (82.74)[10]	14,117 (67.59)[10]
Age (years)				
0-17	N/A	3.42[9]	6656 (1.33)[10]	N/A
18-34	N/A	53.21[9]	187,966 (37.55)[10]	N/A
35-44	N/A	21.89[9]	126,646 (25.30)[10]	N/A
45-64	N/A	19.02[9]	149,487 (29.86)[10]	N/A
≥65	N/A	2.46[9]	29,847 (5.96)[10]	N/A
Ethnicity				
Asian	3.24[9]	5.60[9]	N/A	N/A
Black	0.30[9]	0.30[9]	N/A	N/A
Hispanic	23.50[9]	17.40[9]	N/A	N/A
White	73.00[9]	76.60[9]	N/A	N/A
Region				
Northeast	165,531 (19.83)[11]	2598 (17.86)[11]	73,221 (19.58)[10]	N/A
Midwest	174,620 (20.92)[11]	2393 (16.45)[11]	84,302 (22.55)[10]	N/A
South	313,350 (37.53)[11]	4863 (33.44)[11]	123,556 (33.05)[10]	N/A
West	181,400 (21.73)[11]	4690 (32.25)[11]	92,809 (24.82)[10]	N/A

4.2.2 Identifying Post Contents

From each source, we randomly selected 500 posts. We then manually identified the different categories of shared content for each type of health-related social media. As shown in Table 36, we identified 9 different categories. The first 4 categories were identified for both types of health-related social media (hence, all 4 sources). Of these first 4 categories, 3 were also identified by Lopes and da Silva [60], for example, *share experiences*, which we defined as posts in which a user shared a personal experience related to a health-related topic. This is similar to their *sharing personal experiences*

[9] Based on Sadah et al. [63].
[10] Calculated with user data collected or estimated from this study.
[11] Calculated from user counts reported in the study by Sadah et al. [64].

category, except that we did not restrict our definition to experiences shared in response to another post. *About family* has no equivalent in their scheme, but it can be covered by other categories that they have defined, for example, by asking a specific question about or expressing sadness over a family member's illness. Our share experiences category was also similar to categories in other work, for example, the *personal experience of mental illness* category in the study by Reavley and Pilkington [58], the *personal* category from Lee et al. [59], the *personal event* category from Robillard et al. [78], and the *first-hand experience* category from Alvaro et al. [79]. As Twitter and Google+ are more news-based social media, we identified 5 additional categories from these sources. *Educational material* can be considered equivalent to the *teaching* category defined by Lopes and da Silva [60]. Despite the differences between the categories we defined and those proposed by Lopes and da Silva [60], we believed that our categories are sufficient for a *proof of concept* for automatic post content category classification in the two types of health-related social media that we investigated. It should be noted that the identification of specific experiences is outside the scope of this study; the *share experiences* category is a catch-all for any experiences shared in a health-related post from any source.

Table 36. List of all identified categories for health-related online forums and general social networks.

Category	Health-related online forums	General social networks	Example
Share experiences	Yes	Yes	"I could not work after Tylenol." "I have taken Lipitor every day."
Ask for specific medical advice or information	Yes	Yes	"Is honey allowed for diabetics?"
Request or give psychological support	Yes	Yes	"I hope your diabetes is under control." "We're thinking of you."
About family (not about self)	Yes	Yes	"My son is now nine months old and teething like crazy."
Share news	No	Yes	"Kaiser Permanente Invites Software Developers To Build Apps—Forbes. http://feedly.com/k/Zojwq"
Jokes	No	Yes	"Got any jokes about Sodium Hypobromite? NaBro."
Advertisements	No	Yes	"Check out these two vitamins for one recipe! http://bit.ly/1471dbn"
Personal opinion	No	Yes	"Main frustration of lupus is losing the ability to do things that used to be normal"
Educational material	No	Yes	"Side Effects of Alzheimer's and Dementia Drugs http://bit.ly/cK7L1f"

We asked 3 graduate students to label the selected data from WebMD, Twitter, and Google+; we used a majority vote as the final result for each of these sources. Table 37 lists the intercoder agreement as given by a Krippendorff's alpha for our labeled datasets from WebMD, Twitter, and Google+. The selected DailyStrength data were labeled by the labeler with the highest agreement with the majority averaged over each category from the other 3 sources (average alpha = 0.680). As shown in Table 38, the distribution of categories in each source is different, for example, the share experiences category is more common in health-related online forums (WebMD and DailyStrength).

Table 37. Intercoder agreement for our labeled datasets (Krippendorff's alpha).

Category	WebMD	TwitterHealth	Google+Health
Share experiences	0.349	0.446	0.109
Ask for specific medical advice or information	0.768	0.225	0.108
Request or give psychological support	0.219	0.090	−0.007
About family (not about self)	0.736	0.322	−0.010
Share news	N/A	0.083	0.083
Jokes	N/A	0.177	0.029
Advertisement	N/A	0.220	0.107

Table 38. Percentages of categories in each source from the labeled data[12].

Category	**WebMD, *n* (%)**	**DailyStrength, *n* (%)**	**TwitterHealth, *n* (%)**	**Google+Health, *n* (%)**
Share experiences	236 (47.2)	400 (80.0)	74 (14.8)	65 (13.0)
Ask for specific medical advice or information	270 (54.0)	173 (34.6)	3 (0.6)	10 (2.0)
Request or give psychological support	126 (25.2)	247 (49.4)	9 (1.8)	7 (1.4)
About family (not about self)	68 (13.6)	37 (7.4)	5 (1.0)	34 (6.8)
Share news	N/A	N/A	56 (11.2)	145 (28.9)
Jokes	N/A	N/A	38 (7.6)	33 (6.6)
Advertisement	N/A	N/A	26 (5.2)	70 (14.0)

4.2.3 Bot Filtering

We examined the impact of automated accounts (i.e. *bots*) on our study using OSoMe's Botometer (formerly BotOrNot, Indiana University) [80], a tool that estimates how likely a Twitter account is to be a bot. We used the Botometer API to score each account that has a tweet in our initial sample of 500. The API assigned each of the 345 accounts that were still active a score in the range 0 to 1, with higher scores corresponding to a higher likelihood of an automated account. We manually evaluated each account with a score above 0.5. With this threshold, which was chosen because it is a natural choice that avoids possible bias from a more arbitrary choice of threshold value, we found a total of 33 likely bot accounts. We found that tweets from these accounts make up a substantial portion of the categories share news (11 tweets), advertisement (12 tweets), and educational material (10 tweets). As Botometer's API rate limit makes removing all bot tweets from our Twitter corpus of over 11 million tweets unfeasible, we instead randomly selected 1000 posts from each day in the date range of our Twitter data.

[12] $N = 500$.

For each author of these selected posts, we again used Botometer to evaluate the likelihood of an automated account, removing tweets from accounts with a score above 0.5 for a total of 142,411 tweets used in our analysis.

We also manually examined 100 posts each from WebMD and DailyStrength to determine the prevalence of bots on these websites, which consisted of one of the authors reading each of these posts and determining whether or not it appeared to be posted by a spambot. In the context of online forums, a spambot is an automated agent that posts promotional content [81]. By this criterion, none of the posts examined appeared to have been posted by a bot. Although this does not guarantee that there are no posts from bots in the data from these websites used in our study, it does suggest that posts from bots may be much less prevalent in these sources, likely because of the smaller volume of posts and more active moderation compared with Twitter and Google+.

4.2.4 Building Post Content Classifiers

For each category, we performed binary classification experiments with three classifier algorithms: random forest [82], linear SVM [83], and convolutional neural network (CNN) [84]. We first extracted and concatenated the features shown in Table 39. These features include the title of a post, the main text of a post (body), and the name of the message board that contains the post (board name). For the random forest and SVM classifiers, we converted the features to a term frequency-inverse document frequency vector with stop words removed and the remaining words lemmatized. For the CNN classifier, we converted the features to sets of fastText [85] vectors pretrained on Wikipedia. For all classifiers, we applied class weights to the training data such that the weight of the positive class (the post is in the category) is balanced with the weight of the negative class (the post is not in the category). These weights are used with random forest and SVM according to their implementations by Pedregosa et al. [86], whereas CNN uses oversampling of the least frequent class as recommended by Buda et al. [87].

Table 39. All classifiers' training features.

Source	Extracted features
WebMD	Title, body, and board name
DailyStrength	Title, body, and board name
Google+	Title and body
Twitter	Body

To build the classifiers, we excluded the categories where the percentage is less than 10.0% (50/500), and for the rest, we first split the labeled data to two datasets as follows: (1) a training dataset (450 posts) and (2) a test dataset (50 posts), held out for a final test after training is complete. Afterward, for each classifier algorithm, we trained each classifier by varying the hyperparameters shown in Table 40, considering each combination of hyperparameter values. For all combinations, we performed a 5-fold cross-validation on the training dataset to select the combination of hyperparameter values with the highest balanced accuracy [88]. Finally, we used these hyperparameter values to create a model trained on the full training dataset and tested this model on the test dataset that was held out before the cross-validation experiments. Note that we did not use a nested cross-validation, as our goal in these experiments was to find a single combination of hyperparameter values that we could use to apply a sufficiently accurate classifier model to the rest of our data.

Table 40. Classifier hyperparameter values evaluated in our experiments.

Classifier and hyperparameter	Values
Random forest	
Maximum tree depth	2, 4, 8, 16, 32, 64
Number of trees	10, 100, 1000
Support vector machine	
C	0.001, 0.01, 0.1, 1, 10
Loss function	Hinge, squared hinge
Convolutional neural network	
Filter window sizes	(2, 3, 4), (3, 4, 5), (4, 5, 6)
Feature maps per filter window size	100, 200, 300, 400, 500, 600

Table 41 shows the classifiers' accuracy for WebMD, DailyStrength, Twitter, and Google+. We have shown only the classifiers for categories that have more than 10% of labeled data.

Table 41. Classifier results for each category[13].

Source and category	Random forest		SVM		CNN	
	Accuracy	**Balanced accuracy**	**Accuracy**	**Balanced accuracy**	**Accuracy**	**Balanced accuracy**
WebMD						
Share experiences	82%	*0.83*	82%	0.81	82%	0.82
Ask for advice	80%	0.82	82%	*0.83*	74%	0.76
Psychological support	78%	0.71	86%	*0.8*	76%	0.68
About family	76%	0.56	80%	*0.89*	94%	0.81
DailyStrength						
Share experiences	82%	0.80	80%	0.70	82%	*0.82*
Ask for advice	78%	0.71	76%	0.70	74%	*0.7*
Psychological support	68%	0.68	66%	0.65	76%	*0.68*
TwitterHealth						
Share experiences	78%	0.77	82%	*0.82*	86%	0.74
Share news	82%	0.64	80%	0.73	94%	*0.81*
Google+Health						
Share experiences	88%	0.48	70%	*0.72*	90%	0.60
Share news	52%	0.48	56%	0.52	66%	*0.59*
Advertisement	76%	0.59	48%	0.53	84%	*0.6*[15]
Personal opinion	78%	0.48	74%	*0.71*	84%	0.60
Educational material	80%	0.66	68%	0.76	82%	*0.79*

[13] N = 50. The category of each source-category combination with at least one classifier that achieved a balanced accuracy of at least 0.75 is italicized for emphasis, as is the highest balanced accuracy for each source-category combination.

For the remainder of our analysis, we only considered source-category combinations with a classifier that achieved a balanced accuracy higher than 0.75.

For the source-category combinations that did not have a classifier that achieved a balanced accuracy of at least 0.75, we performed another round of experiments in which we attempted to classify posts using the best-performing classifier trained on a corresponding category from another source, for example, random forest for share experiences from WebMD. In these experiments, we used 500 posts from one source for training and 500 posts from another source for testing and again finding the best combination of hyperparameters via a 5-fold cross-validation of the training data. Table 42 shows the results of these experiments. Classifiers trained on the DailyStrength and Twitter data achieved a balanced accuracy of over 0.75 on the share experiences category from Google+, so we added this category to the set of categories considered for further analysis. For each category in this set, we used the model with the highest balanced accuracy for that category to label the rest of the data. We reported our findings on the frequency of these categories by several demographics according to their respective classifiers in the Results section.

Table 42. Results of classifiers trained on a corresponding category from another source[14].

Training source	**Test source**	**Category**	**Classifier**	**Accuracy**	**Balanced accuracy**
WebMD	DailyStrength	Psychological support	SVM	65.6%	0.656
WebMD	Google+Health	Share experiences	Random forest	85.6%	0.584
DailyStrength	*Google+Health*	*Share experiences*	CNN	76.6%	*0.800*
Twitter	*Google+Health*	*Share experiences*	SVM	81.6%	*0.770*
Twitter	Google+Health	Share news	CNN	72.0%	0.562

4.2.5 Demographic Analysis

We chose four demographic attributes as shown in Table 34: gender, age, ethnicity, and location. Where possible, we extracted these attributes from user profiles. These attributes are not available for every source, so we used existing classifier models where available to estimate their values. Specifically, we used the classifiers from Mislove et al. [68] to estimate gender for Twitter users and ethnicity for both Twitter and Google+ users. To estimate gender for WebMD users, we used the classifier from Sadah et al. [63], an extension of the classifier by Mislove et al. that considers a user's screen name when the user's first name is not present. These classifiers use the 1000 most popular male and female birth names reported by the US Social Security Administration for each year from 1935 to 1995 as ground truth for gender and the distribution of ethnicities for each last name as reported by the 2000 US Census as ground truth for ethnicity. For each of these attributes, we used the data labeled by our post content category classifiers to determine how frequently users of each demographic write a post with one of these categories, for example, the percentage of posts made by male users in which a user

[14] N = 500. The test source, category, and balanced accuracy of each classifier that achieved a balanced accuracy of at least 0.75 are italicized for emphasis.

shared his experiences. When comparing these percentages, we calculated statistical significance via a Pearson chi-square test. Note that a post can be in more than one category, for example, a post can both share experiences and ask for medical advice.

4.2.6 Top Distinctive Message Boards

For each combination of demographic and category (e.g. male and share experiences) analyzed in WebMD and DailyStrength, we found the most distinctive message boards for that combination. For WebMD, we considered only boards that have at least 0.01% of posts for a given combination, or 30 if 0.01% is less than 30. Owing to the large number of message boards on DailyStrength (1608 analyzed in this study), we reduced this restriction to only consider boards with at least 30 posts for a given combination. We then determined distinctiveness by calculating the relative difference of each board. On the basis of the calculation for top distinctive terms by Sadah et al. [64], we calculated the relative difference of board *b* within the combination of category *c* and demographic *d* of demographic attribute *a* as shown in Equation 1:

$$RelDif_{cd}(b) = \frac{Freq_{cd}(b) - AvgFreq_{ca}(b)}{AvgFreq_{ca}(b)} \qquad (1)$$

where $Freq_{cd}(b)$ is the normalized frequency of posts on board *b* in category *c* by a user in demographic *d*, for example, the number of posts on the WebMD Breast Cancer message board that share experiences and were written by a female user divided by the number of posts on WebMD that share experiences and were written by a female user. $AvgFreq_{ca}(b)$ is the average $Freq_{cd}(b)$ across all demographics *d* within the demographic attribute *a*, for example, male and female for the demographic attribute gender.

4.3 Results

4.3.1 Demographics

In this section, we present the categories' results by each demographic where possible. For age demographics, we organized users into five groups: 0 to 17 years, 18 to 34 years, 35 to 44 years, 46 to 64 years, and older than 65 years. For ethnicity, we considered four possibilities: Asian, black, Hispanic, and white. For location, we considered the four regions designated by the US Census Bureau: Midwest, Northeast, South, and West. As explained in the Methods section, we considered the following categories for each source: (1) WebMD: share experiences, ask for advice, psychological support, and about family; (2) DailyStrength: share experiences and ask for advice; (3) TwitterHealth: share experiences and share news; and (4) Google+Health: share experiences and educational material.

4.3.2 WebMD

As shown in Table 34, our WebMD dataset includes gender predicted by the gender classifier from Sadah et al. [63]. Therefore, we have reported the distribution of gender among its categories. Table 43 shows the frequency of posts made by male and female users for each category. We found that 70.04% (4741/6769) of posts written by male WebMD users asked for advice, compared with 45.14% (6372/14,117) of posts by female users ($p < 0.001$). Table 44 shows the top 10 most distinctive WebMD message boards by the number of posts for each combination of gender and category. Unsurprisingly, these results show that female users were more likely to post on boards about pregnancy and parenting than males in all categories, whereas male users were more likely to discuss men's health issues. Men also gave psychological support and discussed family members on the message board for the infertility drug, Clomid, more frequently than women.

Table 43. WebMD category frequency by gender.

Category	**Gender, *n* (%)**	
	Male (*n* = 6769)	**Female (*n* = 14,117)**
Share experiences	3290 (48.60)	4835 (34.25)
Ask for advice	4741 (70.04)	6372 (45.14)
Psychological support	1914 (28.28)	5515 (39.07)
About family	1986 (29.34)	3623 (25.66)

Table 44. Top 10 most distinctive WebMD message boards for male and female users in each category.

Gender	Share experiences	Ask for advice	Psychological support	About family
Male	• Men's Health • Erectile Dysfunction • Relationships and Coping • Cholesterol Management • Epilepsy • Depression • Allergies • Oral Health • Knee & Hip Replacement • Ear, Nose & Throat	• Erectile Dysfunction • Cholesterol Management • Men's Health • HIV/AIDS • Depression • Epilepsy • Prostate Cancer • Sports Medicine • Pain Management • Ear, Nose & Throat	• Relationships and Coping • Epilepsy • Depression • Back Pain • Heart Disease • Pain Management • Anxiety & Panic • Clomid • Diabetes • Parenting: 4 & 5-Year-Olds	• Relationships and Coping • Depression • Erectile Dysfunction • Back Pain • Clomid • Epilepsy • Anxiety & Panic • Pain Management • Sleep Disorders • Digestive Disorders
Female	• Sexual Abuse Survivors Support • Trying to Conceive: 12 Months, Still Trying • Endometriosis • Breast Cancer • Infertility Treatment • Pregnancy: After Infertility • Pregnancy: After 35 • Parenting: Elementary Ages • Self-Harm • Menopause	• Trying to Conceive: 12 Months, Still Trying • Infertility Treatment • Dieting Club: 25-50 Lbs • Parenting: Preteens & Teenagers • Skin & Beauty • Breast Cancer • Food & Cooking • Lupus • Parenting: 3-Year-Olds • Parenting: 9-12 Months	• Chronic Fatigue Syndrome • Lupus • Sexual Abuse Survivors Support • Breast Cancer • Endometriosis • Dieting Club: 10-25 Lbs • Trying to Conceive: 12 Months, Still Trying • Pregnancy: After 35 • Dieting Club: 100+ Lbs • Pregnancy: After Infertility	• Sexual Abuse Survivors Support • Pregnancy: After 35 • Trying to Conceive: 12 Months, Still Trying • Trying to Conceive: After Loss • Breast Cancer • Self-Harm • Parenting: Preteens & Teenagers • Parenting: 9-12 Months • Dieting Club: 50-100 Lbs • Parenting: 6-9 Months

4.3.3 DailyStrength

For our DailyStrength demographic attributes, gender, age, and location, we reported the results for the categories share experiences and ask for advice. Table 45 shows the category frequencies for each demographic. The majority of posts (over 80%) from every demographic share experiences; but among the different age demographics, we saw a clear decline in frequency as age increases, from 92.77% (6175/6656) for users aged younger than 18 years to 81.82% (24,420/29,847) for users 65 years and older (*P*<.001).

The frequency of posts that ask for advice is similar for almost every demographic (30%-40%), with the exception of posts from users younger than 18 years 25.45% (1694/6656). For all comparisons between users younger than 18 years and other age groups, $p < 0.001$.

Table 45. DailyStrength category frequency by gender, age, and location.

Attribute and demographic	**Total number of participants**	**Share experiences, *n* (%)**	**Ask for advice, *n* (%)**
Gender			
Male	95,269	78,760 (82.67)	31,706 (33.28)
Female	456,600	409,640 (89.72)	167,867 (36.76)
Age group (years)			
0-17	6656	6175 (92.77)	1694 (25.45)
18-34	187,966	173,226 (92.16)	65,191 (34.68)
35-44	126,646	113,796 (89.85)	48,335 (38.17)
45-64	149,487	127,089 (85.02)	54,008 (36.13)
≥65	29,847	24,420 (81.82)	10,581 (35.45)
Region			
Northeast	73,221	65,761 (89.81)	28,196 (38.51)
Midwest	123,556	76,630 (90.90)	31,600 (37.48)
South	123,556	110,597 (89.51)	46,933 (37.99)
West	92,809	76,797 (82.75)	31,481 (33.92)

Tables 46-48 show the top 10 most distinctive DailyStrength message boards by the number of posts for each combination of gender and category, age group and category, and location and category, respectively. From these lists, we saw a wider variety of topics compared with WebMD, likely because of the large number of message boards on DailyStrength. However, we still saw some trends when considering broader topics. Male users tend to share experiences on message boards related to personal and social issues. Both male and female users asked for advice most frequently on boards related to physical conditions.

Table 46. Top 10 most distinctive DailyStrength message boards for male and female users in each category.

Gender	Share experiences	Ask for advice
Male	• Vow To Live LGBT Against Suicide • Christian Church 24.7 Ministry • Gay Men's Challenges • Single Dads • GOYA • Dealing with Diabetes2 and remembering Goldi • A Child Abuse Survivors Group • CALM and EASY GAMES • Financial Challenges • Liars Anonymous	• A Laughter Club • Dealing with Diabetes2 and remembering Goldi • Impotence & Erectile Dysfunction • Sex/Pornography Addiction • High Cholesterol • Tinnitus, Deafness and Ear Problems • Urinary Incontinence • Atrial Fibrillation (AFib) • MRSA • LDN .. Low Dose Naltrexone
Female	• helping with the housework • Lesbian Relationship Challenges • prompts • AlAnon One Day At A Time • Daughters of Abusive Mothers • Breastfeeding • Parenting Toddlers (1-3) • Post-Partum Depression • Infertility • Vulvar Cancer	• Pregnancy • Menopause • Trying To Conceive • Miscarriage • Polycystic Ovarian Syndrome (PCOS) • Family & Friends of Bipolar • WHY WEIGHT? LET'S LOSE WEIGHT AND FEEL GREAT! • Infertility • Vulvar Cancer • Breastfeeding

Table 47. Top 10 most distinctive DailyStrength message boards for each age group in each category.

Age group (years)	Share experiences	Ask for advice
0-17	• Weight Loss For Teens • Gay & Lesbian Teens • Depression–Teen • Bipolar Disorder–Teen • Self-Injury • Transgender • Depression • Coming Out • Bisexuality • Eating Disorders	• Weight Loss For Teens • Depression–Teen • Self-Injury • Eating Disorders • Anxiety
18-34	• Sunny and Peaceful Skies • Parenting Toddlers (1-3) • Daily Positive Thoughts • Trying To Conceive • Parenting Newborns & Infants (0-1) • College Stress • Arnold-Chiari Malformation • ALL MOODY BLUES • Career Changes • Cerebral Palsy	• Trying To Conceive • Neuropathy • Pregnancy • Miscarriage • Polycystic Ovarian Syndrome (PCOS) • Cerebral Palsy • Endometriosis • Pseudotumor Cerebri • Sexually Transmitted Diseases–Female • Schizophrenia
35-44	• Vow To Live LGBT Against Suicide • Parenting ‘Tweens (9-12) • Twins, Triplets & More • Self-Hate Syndrome • Parents Whose children have been sexually abused • HOPEFUL HEARTS...LIVING AGAIN AFTER THE LOSS • Neurofibromatosis • Breastfeeding • Hyperparathyroidism • Stillbirth	• kindredspirits • Hyperparathyroidism • Multiple Sclerosis (MS) • Pseudotumor Cerebri • Allergies • Hemochromatosis • Hypothyroidism • Addison’s Disease • MCTD • Graves’ Disease
45-64	• acoa sanctuary • prompts • Christians with MS • InHisCare Bible Study • The Serenity Room • Ticked off about Lyme • Biblical Studies and Archaeology • Alanon support group • Just support • WHY WEIGHT? LET’S LOSE WEIGHT AND FEEL GREAT!	• WHY WEIGHT? LETS LOSE WEIGHT AND FEEL GREAT! • MS People Dealing with MS Pain • Dealing with Diabetes2 and remembering Goldi • Multiple Myeloma • Menopause • High Cholesterol • LDN .. Low Dose Naltrexone • Myofascial Pain Syndrome • Neurocardiogenic Syncope • Amputees
≥65	• Banana • A Little Bit Of Kindness Goes A long Way! • AlAnon One Day At A Time • VOICES OF RECOVERY • The Walking Group • The Front Porch • Over The Fence • Muscular Dystrophies • CALM and EASY GAMES • movie lovers	• AlAnon One Day At A Time • VOICES OF RECOVERY • I can’t HEAR you! • COPD & Emphysema • Meniere’s Disease • Parkinson’s Disease • Sleep Apnea • Interstitial Cystitis (IC) • Atrial Fibrillation (AFib) • Acromegaly

Table 48. Top 10 most distinctive DailyStrength message boards for each region in each category.

Region	Share experiences	Ask for advice
Northeast	• WHY WEIGHT? LET’S LOSE WEIGHT AND FEEL GREAT! • Self-Hate Syndrome • Smoking Addiction & Recovery • Urinary Incontinence • Families of Prisoners • Agoraphobia & Social Anxiety • Cocaine Addiction & Recovery • Obesity • CHRISTIAN PARENTS of ESTRANGED ADULT CHILDREN • Brain Injury	• WHY WEIGHT? LET’S LOSE WEIGHT AND FEEL GREAT! • Obesity • Hidradenitis Suppurativa • Endometriosis • Deep Vein Thrombosis (DVT) • Atrial Fibrillation (AFib) • Diets & Weight Maintenance • Gastritis • Polycystic Kidney Disease (PKD) • Hypothyroidism
Midwest	• Just support • acoa sanctuary • helping with the housework • kindredspirits • The Coffee Shop • aa Spoken Here • Highly Sensitive People HSP • Financial Challenges • I can’t HEAR you! • Pseudotumor Cerebri	• kindredspirits • Neurocardiogenic Syncope • Pseudotumor Cerebri • Gastritis • Irritable Bowel Syndrome (IBS) • COPD & Emphysema • Parkinson’s Disease • Polycystic Kidney Disease (PKD) • Pancreatitis • Graves’ Disease
South	• prompts • Beyond Medication • InHisCare Bible Study • Ticked off about Lyme • Muscular Dystrophies • aa friends • Anxiety and POSITIVE CHOICES • Games for Fun and Relaxation • MS People Dealing with MS Pain • Parents Whose children have been sexually abused	• MS People Dealing with MS Pain • High Cholesterol • Cirrhosis • Polymyositis & Dermatomyositis • Addison’s Disease • Meniere’s Disease • MCTD • Trying To Conceive • Endometriosis • Polycystic Ovarian Syndrome (PCOS)
West	• A Little Bit Of Kindness Goes A long Way! • The Walking Group • Alanon support group • VOICES OF RECOVERY • AlAnon One Day At A Time • BIBLICAL STUDIES • The Sunflower group • My Favorite Things. • FrIeNdShIpRoOm • three prayerpraise	• AlAnon One Day At A Time • Banana • The Sunflower group • WINGS • VOICES OF RECOVERY • A Laughter Club • FrIeNdShIpRoOm • Myofascial Pain Syndrome • Hemochromatosis • Colon Cancer

4.3.4 Twitter

For our Twitter demographic attributes, gender, ethnicity, and location, with gender and ethnicity predicted by the classifier from Mislove et al. [68], we reported the results for categories share experiences and share news using our sample of 142,411 tweets in Table 49. As described in the Methods section, this dataset was created from our full corpus by first sampling 1000 posts for each day represented in the dataset and then pruning tweets from likely bot accounts. All demographics analyzed shared experiences more often than they shared news. Hispanic users had the largest difference, with 29.16% (826/2833) of them shared experiences versus 5.47% (155/2833) of them shared news ($p < 0.001$). Users from the Northeast census region had the smallest difference, with 20.38% (1093/5362) of them shared experiences versus 10.16% (545/5362) of them shared news; $p < 0.001$. Where comparison is possible between these demographics and their counterparts in WebMD and DailyStrength, we saw that Twitter users shared experiences less frequently ($p < 0.001$ for all such comparisons).

Table 49. Twitter category frequency by gender, ethnicity, and location.

Attribute and demographic	**Total number of participants**	**Share experiences, *n* (%)**	**Share news, *n* (%)**
Gender			
Male	16,092	3188 (19.81)	1277 (7.94)
Female	17,850	4835 (27.09)	1091 (6.11)
Ethnicity			
Asian	626	166 (26.52)	34 (5.43)
Black	56	12 (21)	3 (5)
Hispanic	2833	826 (29.16)	155 (5.47)
White	9992	2259 (22.61)	728 (7.29)
Region			
Northeast	5362	1093 (20.38)	545 (10.16)
Midwest	4686	1084 (23.13)	380 (8.11)
South	9855	2162 (21.94)	850 (8.63)
West	5448	1164 (21.37)	515 (9.45)

We also performed this analysis on our full Twitter dataset of 11,637,888 tweets. We compared these results with the results shown in Table 49 and found that the differences were generally not statistically significant (with statistical significance defined as $p < 0.05$) for the share experiences category but were significant for all but one demographic in the share news category. These findings agree with our evaluation of bot likelihood using our initial sample of 500 tweets, where we found that the share news category had a substantial number of tweets from likely bot accounts, but the share experiences category did not. The *p*-values of these comparisons are shown in Table 50.

Table 50. Significance of comparisons between Twitter results using pruned data and results using all data.

Category	**Male**	**Female**	**Asian**	**Black**	**Hispanic**	**White**	**Northeast**	**Midwest**	**South**	**West**
Share Experiences	< 0.001	0.47	0.24	0.80	0.68	0.15	0.13	0.048	0.002	< 0.001
Share News	< 0.001	< 0.001	< 0.001	0.23	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

4.3.5 Google+

Our Google+ demographic attributes include gender, age, ethnicity, and location, with ethnicity predicted by the classifier from Mislove et al. [68], and for these attributes we reported the results from the share experiences and educational material categories in Table 51. As classifiers trained on our labeled Google+ dataset did not achieve a sufficiently high balanced accuracy for the share experiences category, we considered classifiers trained on the labeled DailyStrength and Twitter data as described in the Methods section. The full set of Google+ posts were classified as 34.13% (63,709/186,666) share experiences by the DailyStrength-trained classifier and 18.83% (35,149/186,666) share experiences by the Twitter-trained classifier. As the latter distribution of the share experiences category is closer to the distribution reported in

Table 38, 13.0% (65/500), we used the Twitter-trained classifier for the remainder of our analysis in the share experiences category.

Table 51. Google+ category frequency by gender, age, ethnicity, and location.

Attribute and demographic	**Total number of participants**	**Share experiences, *n* (%)**	**Educational material, *n* (%)**
Gender			
Male	61,479	15,234 (24.78)	16,200 (26.35)
Female	32,082	9803 (30.56)	8029 (25.03)
Age group (years)			
0-17	42	19 (45.24)	8 (19.05)
18-34	552	189 (34.24)	141 (25.54)
35-44	308	101 (32.79)	46 (14.94)
45-64	499	62 (12.42)	171 (34.27)
≥65	45	9 (20.00)	13 (28.89)
Ethnicity			
Asian	2825	730 (25.84)	1010 (35.75)
Black	72	28 (38.89)	13 (18.06)
Hispanic	3389	1137 (33.55)	707 (20.86)
White	17,230	5076 (29.46)	3340 (19.38)
Region			
Northeast	4510	1097 (24.32)	957 (21.22)
Midwest	4210	1310 (31.12)	716 (17.01)
South	9532	2636 (27.65)	1913 (20.07)
West	7959	2279 (28.63)	1708 (21.46)

From these results, we saw that most demographics appeared to share experiences more frequently than the set of all Google+ users. This is likely the effect of a bias toward users who chose to report these attributes (or a real name, in the case of ethnicity). When comparing how often a demographic shares experiences with how often posts from users with no data on that demographic’s corresponding attribute share experiences (e.g. posts from men vs. posts from users who did not report gender), we found that $p < 0.001$ for all such comparisons except for users aged ≥ 65 years ($p = 0.83$). Where comparison is possible between these demographics and their counterparts in WebMD and DailyStrength, we saw that Google+ users shared experiences less frequently ($p < 0.001$ for all such comparisons).

Educational material was shared less frequently by users aged between 35 and 44 years, 14.9% (46/308) than by users of any other age group. In particular, they shared educational material much less frequently than both the previous age group, 18 to 34 years, 25.5% (141/552), $p < 0.001$; and the following age group, 45 to 64 years, 34.3% (171/499), $p < 0.001$. Asian Google+ users, 35.75% (1010/2825), substantially shared more educational material than users of any other ethnicity ($p = 0.002$ vs. black users, $p < 0.001$ vs. Hispanic users, and $p < 0.001$ vs. white users).

4.4 Discussion

4.4.1 Principal Findings

Our analysis shows several interesting results. From our initial samples, we found that health-related posts from general social networks often shared news and educational material, and posts on health-related online forums frequently shared experiences, asked for medical advice, and requested or gave psychological support (Table 38). Our evaluation of three classification algorithms on the post content categories described by our study showed that, in terms of balanced accuracy, SVM tended to perform well on WebMD, whereas CNN performed better on DailyStrength data. Of the 2 Twitter categories used in our experiments, share experiences and share news, SVM performed the best in share experiences and CNN was the best in share news. None of the classifiers we evaluated performed particularly well when trained with the Google+ data; only the CNN classifier was able to meet our performance threshold in the Google+ educational material category. However, in the share experiences category, classifiers trained on the DailyStrength and Twitter data were able to meet our performance threshold in the

Google+ share experiences category, suggesting that at least some transferability is possible with classifiers trained on other datasets.

A further analysis of our health-related online forum data showed distinct differences between users of WebMD and DailyStrength. On WebMD, we found that the majority of posts made by male users and almost half of all posts made by female users asked for advice. This would seem to contradict an earlier study that found that women were the predominant users of the internet for health advice [89], but when considering the overall number of posts from male and female WebMD users included in our study (41,422 posts by men vs. 93,293 by women), we saw that posts asking for advice were still more likely to be written by a woman than a man. DailyStrength users shared experiences frequently in all demographics analyzed in our study, even more so than WebMD users; however, asking for advice was less common than on WebMD. These differences may be explained by the differences in the 2 health-related online forums; although DailyStrength offers support groups for a variety of topics, WebMD communities are often frequented by experts who can provide advice to users.

An analysis of health-related posts on general social networks, Twitter and Google+, suggested differences that they have from health-related online forums. Compared with WebMD and DailyStrength, sharing experiences, which identifies posts in which a user shared a personal experience related to a health-related topic, is far less frequent in posts from Twitter and Google+ that contain one or more of the health-related keywords used in this study. The relatively low frequency of sharing experiences in our sample of several health-related topics on general social networks compared with the frequency of sharing experiences on health-related online forums may be due to a variety of factors, such as Twitter's lack of health-related communities because of its structure as well as WebMD's and DailyStrength's focus on answering medical questions and providing support, respectively. Some subsets of health-related tweets studied in other work have low proportions of sharing experiences similar to our observations, such as tweets about depression [58], schizophrenia [58], and dementia [78], as well as tweets from health-related Twitter users [59]. However, other work has shown that the proportion can be much higher, such as in tweets about dental pain [90] and prescription drug use [79]. Many health-related topics had high proportions of posts that shared experiences in our Google+ data, for example, *headache*, 93.22% (6572/7050); *migraine*, 78.77% (2029/2576); *insomnia*, 71.41% (2430/3403); *cold sore*, 58.0% (370/638); and *diazepam*, 51.1% (95/186). This suggests that the proportion of sharing experiences in health-related posts may be highly dependent on the topic or topics studied; thus, our findings on the share experiences category may not generalize to other studies on health-related social media posts.

Our comparison of results between our stratified sample of Twitter data with tweets from suspected bots removed and our full Twitter dataset showed that automated accounts had a significant impact on the share news category. Other work has also shown that bots can have an effect on health-related Twitter conversations, particularly on the subject of vaccination. Bots post both pro- and antivaccine tweets [91] and retweet vaccine-related tweets at higher frequencies than human users [92]. The use of bots in this manner amplifies the debate and further polarizes the communities involved. It is clear that bot activity must be considered when analyzing health-related conversations on Twitter.

The differences in how often educational material is shared on Google+ between the demographics we studied highlight potential targets for informational health care campaigns. A health care campaign is a health care–related broad nationally or subnationally driven, led, or coordinated activity [93]. Users in the age demographic of 35 to 44 years, who share educational material less often than other age groups, may benefit from being provided with medical information that they are not aware of. Demographics that share educational material more frequently than others, such as Asian Google+ users, may also be of interest to medical experts. If a further analysis of the educational material shared by these groups shows that the information is inaccurate or misleading, providing correct information may benefit them.

Our results provide useful information that can help health care providers to reach the right demographic group. For example, researchers looking for clinical trial participants can use health-related online forums, where many posts are about sharing experiences. Moreover, demographic-specific results can help guide the targeted educational campaigns. As an example, male WebMD users ask specific medical advice questions more often than females, so male WebMD users may be more receptive to a campaign offering advice from medical experts.

The classifier models used in this study can also be useful for researchers who want to study posts that contain the categories we studied. For example, a researcher who wants to study experiences about a particular drug can use these classifiers to find posts that share experiences from a larger dataset of posts that mention that drug. As another example, a researcher who wants to find out which disorders are frequently mentioned among users who share news can use a classifier to gather a dataset of news-sharing posts. In general, we provided researchers with tools that enable them to answer hypotheses and do research on the subject of health-related social media posts. These tools are provided by the description of our methodology, which describes how one might build these classifier models, and by trained classifier models that are available on request. Similar tools may also be applicable to the categories in the scheme proposed by Lopes and da Silva [60]. We leave this as future work.

4.4.2 Limitations

As users of health-related social media use an informal writing style, our selected 274 words to filter Twitter and Google+ as described in the Methods section may not cover all health-related posts or their variability in topics. For example, the abbreviation *IUI* (intrauterine insemination) is widely used in health-related posts but not included in the health-related keyword list. Another limitation is the different uses of terms used to filter Twitter and Google+. For example, the word “cancer” yields many tweets that talk about zodiac signs.

We found that some Twitter categories have a high proportion of tweets from automated accounts. Although we have attempted to filter out tweets from such accounts, some such tweets may still exist in the data used in our analysis, and tweets from legitimate accounts may have been filtered out. Our initial evaluation of bot prevalence also found that the educational material category had a high proportion of tweets from bots. This may be also true of that category in the Google+ data, which was not filtered for bots; thus, those results may not accurately represent the demographics studied.

Our demographic populations may not be fully representative of all users from the sources in our study. As shown in Table 34, some of our demographics were estimated using classifiers, and these estimates are not always correct. Other demographics in our study are optionally reported by users. This introduces a bias toward users who choose to report their age, gender, and/or location, as noted in our results from Google+. We also assumed these reported demographics are correct for each such user.

4.4.3 Conclusions

In this study, we analyzed the content shared in two different types of health-related social media: health-related online forums and general social networks. For the two types of health-related social media, we manually identified 4 post categories: share experiences, ask for specific medical advice, request or give psychological support, and about family; and we additionally identified 5 categories for general social networks: share news, jokes, advertisements, personal opinion, and educational material. After labeling randomly selected data for each source, we built classifiers for each category. Finally, we made demographic-based content analyses where possible.

Chapter 5

Automatic Classification of Online Doctor Reviews: Evaluation of Text Classifier Algorithms

Background: An increasing number of doctor reviews are being generated by patients on the internet. These reviews address a diverse set of topics (features), including wait time, office staff, doctor's skills, and bedside manners. Most previous work on automatic analysis of Web-based customer reviews assumes that (1) product features are described unambiguously by a small number of keywords, for example, *battery* for phones and (2) the opinion for each feature has a positive or negative sentiment. However, in the domain of doctor reviews, this setting is too restrictive: a feature such as *visit duration* for doctor reviews may be expressed in many ways and does not necessarily have a positive or negative sentiment.

Objective: This study aimed to adapt existing and propose novel text classification methods on the domain of doctor reviews. These methods are evaluated on their accuracy to classify a diverse set of doctor review features.

Methods: We first manually examined a large number of reviews to extract a set of features that are frequently mentioned in the reviews. Then we proposed a new algorithm that goes beyond bag-of-words or deep learning classification techniques by leveraging natural language processing (NLP) tools. Specifically, our algorithm automatically extracts dependency tree patterns and uses them to classify review sentences.

Results: We evaluated several state-of-the-art text classification algorithms as well as our dependency tree–based classifier algorithm on a real-world doctor review dataset. We showed that methods using deep learning or NLP techniques tend to outperform traditional bag-of-words methods. In our experiments, the 2 best methods used NLP techniques; on average, our proposed classifier performed 2.19% better than an existing NLP-based method, but many of its predictions of specific opinions were incorrect.

Conclusions: We conclude that it is feasible to classify doctor reviews. Automatically classifying these reviews would allow patients to easily search for doctors based on their personal preference criteria.

5.1 Introduction

5.1.1 Background

The problem of automatic reviews analysis and classification has attracted much attention because of its importance in ecommerce applications [94-96]. Recently, there has been an increase in the number of sites where users rate doctors. Several works have analyzed the content and scores of such reviews, mostly by examining a subset of them through qualitative and quantitative analysis [21, 97-101] or by applying text-mining techniques to characterize trends [102-104]. However, not much work has studied how to automatically classify doctor reviews.

In this study, our objective was to automatically summarize the content of a textual doctor review by extracting the features it mentions and the opinion of the reviewer for each of these features; for example, to estimate if the reviewer believes that the wait time or the visit time is long or if the doctor is in favor of complementary medicine methods.

We explore the feasibility of reaching this objective by defining a broader definition of the review classification problem that addresses challenges in the domain of doctor reviews and examining the performance of several machine learning algorithms in classifying doctor review sentences.

Previous work on customer review analysis focused on automated extraction of features and the polarity (also referred as opinion or sentiment) of statements about those features [95, 105, 106]. Specifically, these works tackle the problem in 2 steps: first they extract the features using rules, and then, for each feature, they estimate the polarity using hand-crafted rules or supervised machine learning methods. This works well if (1) the features are *basic*, such as the battery of a phone, which are generally described by a single keyword, for example, *the battery of the camera is poor*, and (2) the opinion is objectively positive or negative but does not support more subjective features like visit time, where for some patients it is positive to be longer, and for some, it is negative. In other words, statements about features in product reviews tend to be more straightforward and unambiguously positive or negative, whereas reviews on service, such as doctor reviews, are often less so, as there may be many ways to express an opinion on some aspect of the service.

In our study, the features may be more complex, for example, the *visit time* feature can be expressed by different phrases such as “spends time with me,” “takes his time,” “not rushed,” and so on. As another example, “appointment scheduling” can be expressed in many different ways, for example, “I was able to schedule a visit within days” or “The earliest appointment I could make is in a month." Other complex classes include *staff* or *medical skills*.

Furthermore, in our study, what is positive for one user may be negative for another. For example, consider the sentence "Dr. Chan is very fast so there is practically no wait time and you are in and out within 20 minutes." The sentiment in this sentence is positive, but a short visit implied by *in and out within 20 minutes* may be negative for some patients. Instead, what we want to measure is long visit time versus short visit time. This is different from work on detecting transition of sentiment [107] because it is not enough to detect the *true* sentiment, but we must also associate it with a class (long visit time vs. short visit time).

To address this variation of the review classification problem, we created a labeled dataset consisting of 5885 sentences from 1017 Web-based doctor reviews. We identified several classes of doctor review opinions and labeled each sentence according to the presence and polarity of these opinion classes. Note that our definition of polarity is broader than in previous work as it is not strictly positive and negative but rather takes the subjectivity of patient opinions into account (e.g. complementary medicine is considered good by some and bad by others).

We adapt existing and propose new classifiers to classify doctor reviews. In particular, we consider 3 diverse types of classifiers:

1. Bag-of-words classifiers such as Support Vector Machine (SVM) [108, 83] and Random Forests [82] that leverage the statistical properties of the review text, such as the frequency of each word.

2. Deep learning methods such as Convolutional Neural Network (CNN) [84], which also consider the proximity of the words.
3. Natural Language Processing (NLP)-based classifiers, which leverage the dependency tree of a review sentence [109]. Specifically, we consider an existing NLP-based classifier [110] and propose a new one, the Dependency Tree-Based Classifier (DTC).

DTC generates the dependency tree for each sentence in a review and applies a set of rules to extract dependency tree–matching patterns. These patterns are then ranked by their accuracy on the training set. Finally, the sentences of a new review are classified based on the highest-ranking matching pattern. This is in contrast to the work by Matsumoto et al. [110], which treats dependency tree patterns as features in an SVM classifier.

The results of our study show that classifying doctor reviews to identify patient opinions is feasible. The results also show that DTC generally outperforms all other implemented text classification techniques.

Here is a summary of our contributions:

1. We propose a broader definition for the review classification problem in the domain of doctor reviews, where the features can be complex entities and the polarity is not strictly positive or negative.
2. We evaluated a diverse set of 5 state-of-the-art classification techniques on a labeled dataset of doctor reviews containing a set of commonly used and useful features.

3. We propose a novel decision tree–based classifier and show that it outperforms the other methods; we have published the code on the Web [111].

5.1.2 Literature Review

In this section, we review research in fields related to this study, which we organize into 5 categories:

- Quantitative and qualitative analysis of doctor review ratings and content
- The application of text mining techniques to describe trends in doctor reviews
- Feature and polarity extraction in customer reviews
- Application of dependency tree patterns to sentiment analysis
- Recent work in text classification

Doctor Review Analysis

Several previous works have analyzed Web-based doctor reviews. Gao et al. described trends in doctor reviews over time to identify which characteristics influence Web-based ratings [21]. They found that obstetricians or gynecologists and long-time graduates were more likely to be reviewed than other physicians, recent graduates, board-certified physicians, highly rated medical school graduates, and doctors without malpractice claims received higher ratings, and reviews were generally positive. Segal et al. compared doctor review statistics with surgeon volume [97]. They found that high-volume surgeons could be differentiated from low-volume surgeons by analyzing the number of numerical ratings, the number of text reviews, the proportion of positive reviews, and the proportion of critical reviews. López et al. performed a qualitative content analysis of doctor reviews [98]. They found that most reviews were positive and identified 3 overarching domains in the reviews they analyzed: interpersonal manner, technical competence, and system issues. Hao analyzed Good Doctor Online, an online health community in China, and found that gynecology-obstetrics-pediatrics doctors were the most likely to be reviewed, internal medicine doctors were less likely to be reviewed, and most reviews were positive [99]. Smith and Lipoff conducted a qualitative analysis of dermatology practice reviews from Yelp and ZocDoc [100]. They found that both the average review scores and the proportion of reviews with 5 out of 5 stars from ZocDoc were higher than those from Yelp. They also found that high-scoring reviews and low-scoring reviews had similar content (e.g. physician competency, staff temperament, and scheduling) but opposite valence. Daskivich et al. analyzed health care provider ratings across several specialties and found that allied health providers (e.g. providers who are neither doctors nor nurses) had higher patient satisfaction scores than physicians, but these scores were also the most skewed [101]. They also concluded that specialty-specific percentile ranks might be necessary for meaningful interpretation of provider ratings by consumers.

Text Mining of Doctor Reviews

Other previous papers have employed text-mining techniques to characterize trends in doctor reviews. Wallace et al. designed a probabilistic generative model to capture latent sentiment across aspects of care [102]. They showed that including their model's output in regression models improves correlations with state-level quality measures. Hao and Zhang used topic modeling to extract common topics among 4 specialties in doctor reviews collected from Good Doctor Online [103]. They identified 4 popular topics across the 4 specialties: the experience of finding doctors, technical skills or bedside manner, patient appreciation, and description of symptoms. Similarly, Hao et al. used topic modeling to compare reviews between Good Doctor Online and the US doctor review website RateMDs [104]. Although they found similar topics between the 2 sites, they also found differences that reflect differences between the 2 countries' health care systems. These works differ from ours in that they use text-mining techniques to analyze doctor reviews in aggregate, while our goal is to identify specific topics in individual reviews.

Customer Review Feature and Polarity Extraction

As discussed earlier in the Introduction, these works operate on a more limited problem setting where the features are usually expressed by a single keyword, and the sentiment is strictly positive or negative. Hu and Liu extracted opinions of features in customer reviews with a 4-step algorithm [95]. This algorithm consists of applying association rule mining to identify features, pruning uninteresting and redundant features, identifying infrequent features, and finally determining semantic orientation of each opinion sentence. Popescu and Etzioni created an unsupervised system for feature and opinion extraction from product reviews [96]. After finding an explicit feature in a sentence, they applied manually crafted extraction rules to the sentence and extracted the heads of potential opinion phrases. This method only works when features are explicit.

Sentiment Analysis with Dependency Trees

There are number of existing works that use dependency trees or patterns for sentiment analysis. A key difference is that our method does not always capture sentiment but the various class labels (e.g. short or long) for each class (e.g. visit time). Hence, we cannot rely on external sentiment training data or on hard-coded sentiment rules, but we must use our own training data.

Agarwal et al. used several hand-crafted rules to extract dependency tree patterns from sentences [112]. They combined this information with the semantic information present in the Massachusetts Institute of Technology Media Lab ConceptNet ontology and employed the extracted concepts to train a machine learning model to learn concept patterns in the text, which were then used to classify documents into positive and negative categories. An important difference from our method is that their dependency patterns generally consist of only 2 words in certain direct relations, while our patterns can contain several more in both direct and indirect relations.

Wawer induced dependency patterns by using target-sentiment (T-S) pairs and recording the dependency paths between T and S words in the dependency tree of sentences in their corpus [113]. These patterns were supplemented with conditional random fields to identify targets of opinion words. In contrast to our patterns, which can represent a subtree of 2 or more words, the patterns in this work are generated from the shortest path between the T and S words.

The work of Matsumoto et al. [110] is the closest to our proposed method, which we experimentally compare in the Results section. They extract frequent word subsequences and dependency subtrees from the training data and use them as features in an SVM sentiment classifier. Their patterns involve frequent words and only include direct relations, whereas our patterns involve high-information gain words and consider indirect relations. Pak and Paroubek follow a similar strategy of extracting dependency tree patterns based on predefined rules and using them as features for an SVM classifier [114]. Matsumoto et al. perform better on the common datasets they considered.

Text Classification

Machine learning algorithms are commonly used for text classification. Kennedy et al. used a random forest classifier to identify harassment in posts from Twitter, Reddit, and The Guardian [115]. Posts were represented through several features such as term frequency-inverse document frequency (TF-IDF) of unigrams, bigrams, and short character sequences; URL and hashtag token counts; source (whether the post was from Twitter); and sentiment polarity. Gambäck and Sikdar used a CNN to classify hate speech in Twitter posts [116]. The CNN model was tested with multiple feature embeddings, including random values and word vectors generated with Word2Vec [117]. Lix et al. used an SVM classifier to determine patient's alcohol use using text in electronic medical records [118]. Unigrams and bigrams in these records were represented using a bag-of-words model.

5.1.3 Problem Definition

Given a text dataset with a set of classes c_1, c_2, …, c_m that represent features previously identified by a domain expert, each class c_i can take 3 values (polarity):

- c_i^0: Neutral. The sentence is not relevant to the class.
- c_i^x, c_i^y: Yes or no. Note that to avoid confusion, we do not say positive or negative, as for some classes such as *visit time* in doctor reviews, some patients prefer when

their visit time is long and some prefer short. In this example, "Yes" could arbitrarily be mapped to *long* and "No" to *short*.

As another example, class c_8 from the doctor review dataset is *wait time* or the time spent waiting to see a doctor. It has 3 possible values: c_8^x, c_8^y, or c_8^0. A sentence with class label c_8^x expresses the opinion that the time spent waiting to see the doctor is short. Examples of c_8^x include "I got right in to see Dr. Watkins," "I've never waited more than five minutes to see him," and "Wait times are very short once you arrive for an appointment." A sentence with class label c_8^y expresses the opinion that the time spent waiting to see the doctor is long. Examples of c_8^y include "There is always over an hour wait even with an appointment," "My biggest beef is with the wait time," and "The wait time was terrible." A sentence with class label c_8^0 makes no mention of wait time. Such sentences may have c_i^x or c_i^y labels from other classes, for example, "This doctor lacks affect and a caring bedside manner" and "His staff, especially his nurse Lucy, go far above what their job requires," or they may instead not be relevant to any class, such as "Dr. Kochar had been my primary care physician for seven years" and "I'll call to reschedule everything." A sentence may take labels from more than one class.

In this study, given a training set T of review sentences with class labels from classes c_1, c_2, …, c_m, we build a classifier for each class c_i to classify new sentences to one of the possible values of c_i. Specifically, we build m training sets T_i corresponding to each class. Each sentence in T_i is assigned a class label c_i^x, c_i^y, or c_i^0.

5.2 Methods

5.2.1 Doctor Reviews Dataset

We crawled Vitals [8], a popular doctor review website, to collect 1,749,870 reviews. Each author read approximately 200 reviews and constructed a list of features. Afterward, through discussions, we merged these lists into a single list of 13 features, which we represent by classes as described in the problem definition (Table 52).

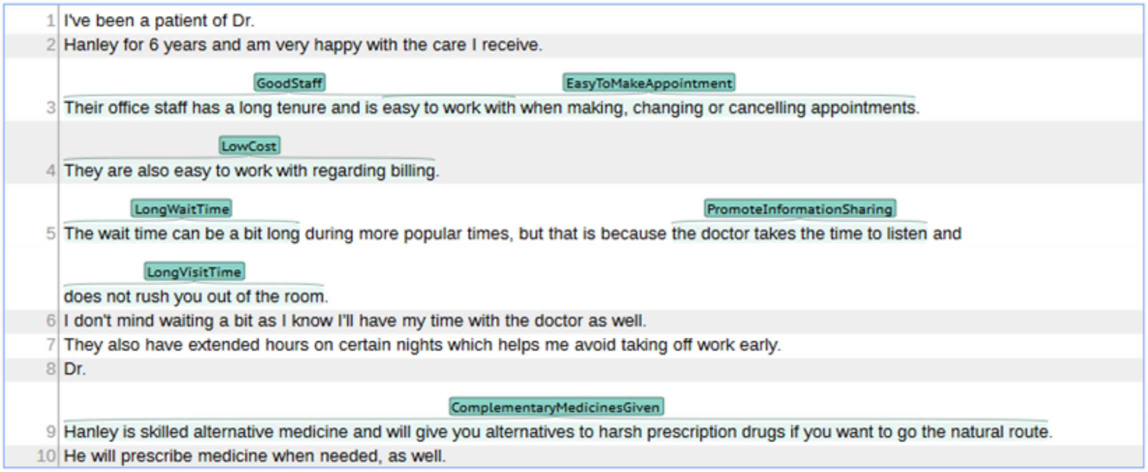
Table 52. Description of initial opinion classes[15].

Class	$\boldsymbol{c_i^x}$	$\boldsymbol{c_i^y}$
Appointment scheduling	Easy to schedule an appointment	Hard to schedule an appointment
Bedside manner	Friendly and caring	Rude and uncaring
Complementary medicine	Promotes complementary medicine	No promotion of complementary medicine
Cost	Inexpensive and billing is simple	Expensive and billing problems
Information sharing	Answers questions and good explanations	Does not answer questions and poor explanations
Joint decision making	Treatment plan accounts for patient opinions	Treatment plan made without patient input
Medical skills	Effective treatment and correct diagnoses	Ineffective treatment and misdiagnoses conditions
Psychological support	Addresses stress and anxiety	Does not address stress and anxiety
Self-management	Encourages active management of care	Does not encourage self-management of care
Staff	Staff is friendly and helpful	Staff is rude and unhelpful
Technology	Uses email, Web-based appointments, and electronic health records	Does not use email and Web-based appointments
Visit time	Spends substantial time with patients	Spends very little time with patients
Wait time	Short time spent waiting to see the doctor	Long time spent waiting to see the doctor

To further filter these classes, we selected 600 random reviews to label. We labeled these reviews using WebAnno, a Web-based annotation tool [119] (Figure 18). Specifically, each sentence was tagged (labeled) with 0 or more classes from Table 52 by

[15] For each class, a sentence that does not mention the class is labeled c_i.

2 of the authors. The union of these labels was used as the set of ground-truth class labels of each sentence; that is, if at least one of the labelers labeled a sentence as c_i^x, that sentence is labeled c_i^x in our dataset.



1 I've been a patient of Dr.
2 Hanley for 6 years and am very happy with the care I receive.
3 Their office staff has a long tenure and is easy to work with when making, changing or cancelling appointments.
4 They are also easy to work with regarding billing.
5 The wait time can be a bit long during more popular times, but that is because the doctor takes the time to listen and does not rush you out of the room.
6 I don't mind waiting a bit as I know I'll have my time with the doctor as well.
7 They also have extended hours on certain nights which helps me avoid taking off work early.
8 Dr.
9 Hanley is skilled alternative medicine and will give you alternatives to harsh prescription drugs if you want to go the natural route.
10 He will prescribe medicine when needed, as well.

Figure 18. Screenshot of WebAnno’s annotation interface with an annotated review.

We found that some of these classes were underrepresented. For each underrepresented class, we used relevant keywords to find and label more reviews from the collected set of reviews, for example, *wait* for wait time and *listen* for information sharing, which resulted in a total of *1017 reviews* (417 in addition to the original 600). These 1017 reviews are our labeled dataset used in our experiments.

Following this, we found that some classes such as complementary medicine and joint decision making were still underrepresented, which we define as having less than 2% of the dataset’s sentences labeled c_i^x or c_i^y, so we omitted them from the dataset. The final dataset consists of 5885 sentences and 8 opinion classes. These classes and the frequency of each of their labels are shown in Table 53.

Table 53. Frequency of each class label in the doctor review dataset.

Class	**Frequency of c_i^x**	**Frequency of c_i^y**	**Frequency of c_i^0**
c_1: appointment scheduling	51	84	5750
c_2: bedside manner	569	341	4975
c_3: cost	25	261	5599
c_4: information sharing	316	136	5433
c_5: medical skills	481	232	5172
c_6: staff	262	368	5255
c_7: visit time	143	79	5663
c_8: wait time	48	199	5638

5.2.2 Background on Dependency Trees

In this section, we describe dependency trees and the semgrex tool that we used for defining matching patterns. Dependency trees capture the grammatical relations between words in a sentence and are produced using a dependency parser and a dependency language. In a dependency tree, each word in a sentence corresponds to a node in the tree and is in one or more syntactic relations between the word or node exactly one other word or node. A dependency tree is a triple $T = \langle N, E, R \rangle$, where

- N is the set of nodes in T where each node $n \in N$ is a tuple containing one or more string attributes describing a word in the sentence T was built from, such as word, lemma, or POS (part of speech)
- E is the set of edges in T where each edge $e \in E$ is a triple $e = \langle n_g, r, n_d \rangle$, where
    - $n_g \in N$ is the governor or parent in relation r
    - r is a syntactic relation between the words represented by n_g and n_d
    - $n_d \in N$ is the dependent or child in relation r
- $R \in N$ is the root node of T

Figure 19 shows a sample dependency tree for the sentence "there are never long wait times." The string representation of this tree, including the parts of speech for its words, is as follows:

[are/VBP expl>there/EX neg>never/RB nsubj>[times/NNS compound>[wait/NN amod>long/JJ]]]

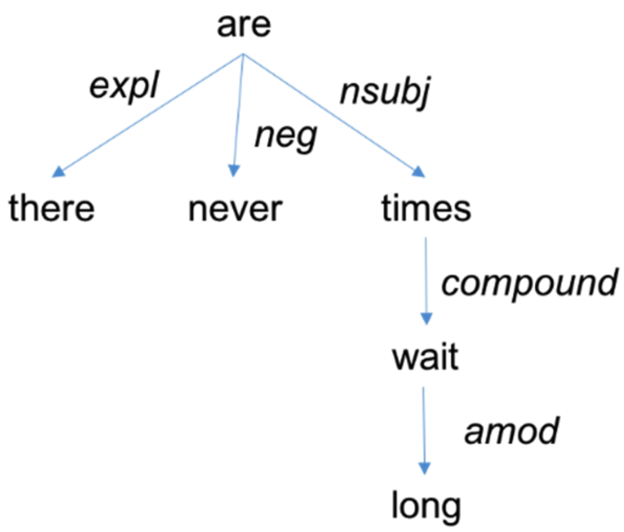


Figure 19. A dependency tree for the sentence "There are never long wait times."

To match patterns against dependency trees, we used Stanford's semgrex utility [120]. In the following, we explain some of the basics of semgrex patterns that help the reader understand patterns presented in this study using descriptions and examples from the Chambers et al. study [120]. Semgrex patterns are composed of nodes and relations between them. Nodes are represented as *{attr1:value1;attr2:value2;…}* where attributes (*attr*) are regular strings such as word, lemma, and pos, and values can be strings or regular expressions marked by "/"s. For example, *{lemma:run;pos:/VB.*/}* means any verb form of the word run. Similar to "." in regular expressions, *{}* means any node in the graph. Relations in a semgrex have 2 parts: the relation symbol, which can be either < or > and optionally the relation type (i.e. *nsubj* and *dobj*). In general, *A<reln B* means *A* is

the dependent of a relation (*reln*) with *B*, whereas *A>reln B* means *A* is the governor of a relation with *B*. Indirect relations can be specified by the symbols >> and <<. For example, *A<<reln B* means there is some node in a dependent→governor chain from *A* that is the dependent of a relation with *B*. Relations can be strung together with or without using the symbol *&*. All relations are relative to first node in string. For example, *A>nsubj B>dobj D* means *A* is a node that is the governor of both an *nsubj* relation with *B* and a *dobj* relation with *D*. Nodes can be grouped with parentheses. For example, *A>nsubj (B>dobj D)* means *A* is the governor of an *nsubj* relation with *B*, whereas *B* is the governor of a *dobj* relation with *D*. A sample pattern that matches the tree in Figure 19 can be:

{} >neg {} >> ({word:wait} > {word:long})

Using the Stanford CoreNLP Java library [121], our proposed classifier builds a dependency tree from a given sentence and determines whether any of a list of semgrex patterns matches any part of the tree.

5.2.3 Proposed Dependency Tree-Based Classifier

Our DTC algorithm is trained on a labeled dataset of sentences as described in the Problem Definition section. On a high level, given a sentence in training dataset T, the classifier generates a dependency tree using the Stanford Neural Network Dependency Parser [122] and extracts semgrex patterns from the dependency tree. These patterns are assigned the same class as the training sentence. When classifying a new sentence, the classifier generates the sentence's dependency tree and assigns a class label to the sentence based on which patterns from the training set match the dependency tree.

In more detail, the classifier's training algorithm generates a sorted list of semgrex patterns, each with an associated class label, from a training dataset T and integer parameters n_i^x, n_i^y, and m. Parameters n_i^x and n_i^y are the maximum number of terms (words or phrases) that will be used to generate patterns of classes c_i^x and c_i^y, respectively. In this study, we only use words, as dependency trees capture relations between words rather than phrases.

The pattern extraction algorithm described in the Pattern Extraction section below receives as input 2 sets W^x and W^y of high-information gain words, for the "Yes" (c_i^x) and "No" (c_i^y) class labels, respectively, from where we pick nodes for the generated patterns. The intuition is that high-information gain words are more likely to allow a pattern to differentiate between the class labels. Considering all words would be computationally too expensive, and it does not offer any significant advantage as we have seen in our experiments. The information gain for W^x is determined by a logical copy of training dataset T in which class labels other than c_i^x are given a new class label $c_i^{x\prime}$, as the words in W^x will be used to identify sentences of class c_i^x. This process is repeated for W^y. Parameter m is the maximum number of these selected words that can be in a single pattern.

The final list of (semgrex pattern p and class label c') pairs is sorted by the weighted accuracy of the pair on the training data, which we define below.

$$\mathrm{WA}(p, T, c_i) = \frac{\sum_{c \in c_i} \mathrm{Accuracy}_c(p,T)}{|c_i|} \qquad (2)$$

We define $\text{Accuracy}_c(p, T)$ as the ratio of training instances in T with class label c that were correctly handled by pattern p. Pattern p, which was paired with class label c', is correct if it matches an instance with class label c' or it does not match an instance without class label c', but it is incorrect if it matches an instance without class label c' or it does not match an instance with class label c'. $|c_i|$ is the number of class labels in class c_i, which is 3 for all of the classes in this study. Intuitively, weighted accuracy treats all class labels with equal importance regardless of their frequency, so patterns that perform well on sentences of often low-frequency class labels c_i^x and c_i^y are assigned higher rank than they would otherwise. The training algorithm is shown in Algorithm 1.

Algorithm 1.

```
train($T, n_i^x, n_i^y, m$):
P := list of semgrex patterns used for classification, initially empty
for all class labels $c$ in $\{c_i^x, c_i^y\}$ do
    $D$ := set of dependency trees for sentences in $T$ with class label $c$
    $T_c$ := copy of $T$ with all non-$c$ class labels given a new class label $c'$
    $W$ := set of top $n_c$ words $w$ in $T_c$ by information gain
    for all trees $t$ in $D$ do
        add all semgrex patterns from extract($W, t, c, m$) to $P$
    end for
end for
test each pattern in $P$ on $T$
sort $P$ by the weighted accuracy of each semgrex pattern tested on $T$ in descending order
return $P$
```

Given a to-be-classified sentence, we compute its dependency tree t and find the highest ranked (pattern p and class label c) pair where p matches t. Then the sentence is classified as c. If no pattern matches the sentence, we provide 2 possibilities: the sentence can be classified as the most common class label in T or it can be classified by a backup classifier trained on T.

5.2.4 Parameters Setting

In all experiments, we use $n_i^x = n_i^y = 30$, as intuitively it is unlikely that there are more than 30 words for a class that can participate in a discriminative semgrex pattern. We set m to 4 for all experiments, because for $m > 4$, it becomes too computationally expensive to compute all patterns.

5.2.5 Pattern Extraction in the Dependency Tree Classifier Algorithm

Overview

Given a dependency tree, we now describe how to extract patterns. Note that we repeat the pattern extraction for the “Yes” and “No” class labels, using W^x and W^y, respectively (W in this section refers to W^x or W^y). We extract semgrex patterns from a dependency tree t with class label c using a set of high-information gain words W and a maximum number of words m. The algorithm returns a set of patterns extracted from t made from up to m words in W.

The rationale for only working with high-information gain words is that we want to generate high-information gain patterns. We also want to preserve negations as they have a great impact to the accuracy of the patterns. If a low information gain word is negated, we replace it by a wildcard (*), which we found to be a good balance for these 2 goals. Each pattern p is associated with c such that a new sentence that matches p is classified as c. Algorithm 2 describes the pattern extraction algorithm.

Algorithm 2.

```
extract($W, t, c, m$):
P := set of patterns, initially empty
$S$ := stack of (tree, word set) pairs, initially empty
**for all** combinations $C$ of words in $W$ with $|C| == min(|W|, m)$ **do**
    $S.\text{push}((t, C))$
**end for**
**while** $S$ is not empty **do**
    $(t', C) := S.\text{pop}()$
    $t' := \text{prune}(t', C)$
    $n$ := root of $t'$
    **while** $n$ = * and $n$ has exactly 1 child **do**
        $n$ := child of $n$
    **end while**
    $t'$ := subtree of $t'$ with root $n$
    remove each "*" node $n'$ in $t'$ with exactly 1 child $c'$, and make the parent of $n'$ the parent of $c'$
    with an *indirect* relation
    add $(\text{pattern}(t'), c)$ to $P$
    **for all** combinations $C'$ of non-* words in $t'$ with $|C'| > 1$ **do**
        $S.\text{push}((t', C'))$
    **end for**
**end while**
**return** $P$
```

```
prune($t, W$):
$t'$ := copy of $t$
recursively prune from $t'$ leaves that do not start with any word in $W$ and are not in a negation
relation
**for all** nodes $n$ in $t'$ **do**
    **if** $n$ does not start with any word in $W$ **then**
        $n$ := *
    **end if**
**end for**
**return** $t'$
```

Details

The algorithm first creates a copy t' of t for each combination C of m words in W and pushes each (t', C) pair onto a stack. For each (t', C) popped from the stack, we execute the following steps:

1. *Create initial subtree*: Prune t' to keep only words in C, negations, and intermediate "*" nodes connecting them.

2. *Remove unimportant nodes*: Eliminate "*" nodes from t' starting with the root if it is a "*" node and has exactly 1 child (the child becomes the new root of t' and this repeats until the root no longer meets these criteria). Subsequently, remove each "*" node n' in t' with exactly 1 child and add an indirect relation edge from the parent of n' to the child of n'.
3. *Add subpatterns*: If $(\text{pattern}(t'), c)$ is not already in P, add $(\text{pattern}(t'), c)$ to the set of patterns P, and then push (t', c') onto the stack for each combination C' of 2 or more non-* words in t'.

The algorithm then moves on to the next item on the stack. Once the stack is empty, we return the resulting set of patterns and their associated class labels.

The $\text{prune}(t, w)$ procedure recursively removes leaf nodes that do not start with any word in W and are not in a negation relation with their parents. Intermediate nodes that connect the remaining nodes and do not start with any word in W are replaced by *. The $\text{pattern}(t)$ procedure converts a dependency tree t to its semgrex format representation. Each "*" node is represented by an empty node *{}*, and most relations are represented by the generic > or >> relations (for direct and indirect relations, respectively), which match any type of relation. An exception to this is the negation relation, which is preserved in the semgrex pattern as the >*neg* token.

Example

Consider a sentence from the doctor review dataset class c_8 (wait time), "I arrived to my appointment on time and waited in his waiting room for over an hour," which has class label c_8^y (long wait). The dependency tree generated from this sentence is shown in Figure 20.

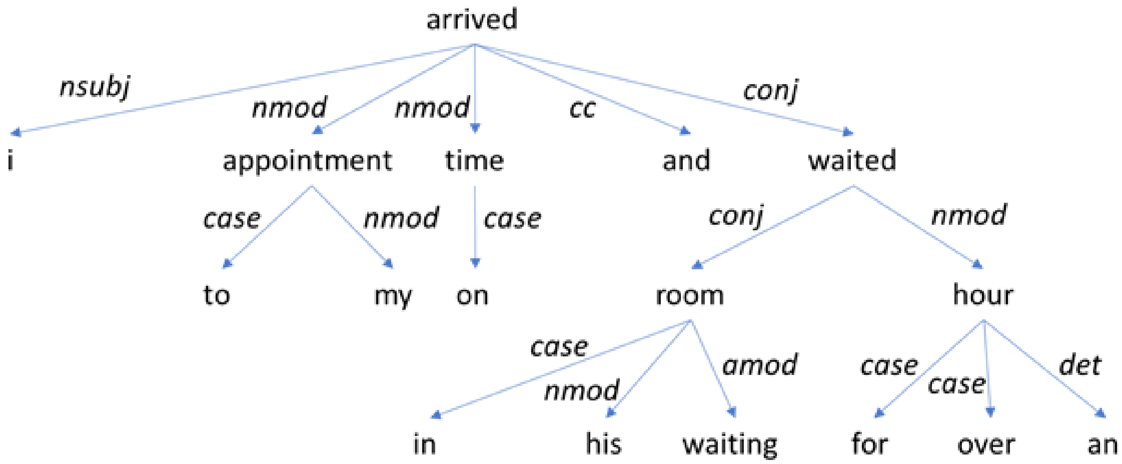


Figure 20. Dependency tree for the sentence “I arrived to my appointment on time and waited in his waiting room for over an hour.”

Among the patterns extracted from this tree are:

1. *{} > {word:/time.*/} >> {word:/hour.*/}*
2. *{word:/arrived.*/} > {word:/time.*/}*
3. *{} > {word:/time.*/} > ({} > {word:/room.*/} > {word:/hour.*/})*
4. *{word:/arrived.*/} >> {word:/hour.*/}*

Pattern 1 means that some node has a direct descendant *time* and an indirect descendant *hour*. Pattern 2 means that *time* is a direct descendant of *arrived*. Pattern 3 means that some node has 2 direct descendants; 1 is *time* and the other is some other node that has direct descendants *room* and *hour*. Finally, pattern 4 means that *hour* is an indirect descendant of *arrived*.

5.3 Results

5.3.1 Classifiers Employed

We consider 3 types of classifiers:

1. Statistical bag-of-words classifiers, which view the documents as bags of keywords:
    - Random Forests (RF): RF, as implemented in Scikit-learn by Pedregosa et al. [86]. Documents are represented with TF-IDF using *n*-grams of 1 to 3 words, a minimum document frequency of 3%, up to 1000 features, stemming, and omission of stop words. The classifier uses 2000 trees. All other parameters are given their default values from [86].
    - SVM: C-support vector classifier as implemented in Scikit-learn by Pedregosa et al. [86], which is based on the implementation from the study by Chang and Lin [123]. Documents are represented with TF-IDF using the same parameters as with random forest. The parameters for the classifier are given their default values from Scikit-learn by Pedregosa et al. [86].
2. Deep learning classifiers:
    - CNN or CNN-W (CNN with Word2Vec): We use 2 variants of the CNN implementation by Britz [124]. Both use the default parameters. The first variant is initialized with a random uniform distribution, as in the CNN implementation by Britz [124]. The second is initialized with values from the Word2Vec model implementation from Gensim by Rehurek and Sojka [125].

- D2V-NN (Doc2Vec Nearest Neighbor): A nearest neighbor classifier that uses the Doc2Vec model [126] implementation from Gensim by Rehurek and Sojka [125]. Documents are converted to paragraph vectors and classified according to the nearest neighbor using cosine similarity as the distance function.

For CNN-W and D2V-NN, the Word2Vec and Doc2Vec models, respectively, are trained on an unlabeled set of 8,977,322 sentences from the collected doctor reviews that were not used to create the labeled dataset.

3. NLP classifiers, which exploit the dependency trees of a review's sentences:
   - Matsumoto: We implemented the method described in the study by Matsumoto et al. [110] using the best-performing combination of features from their experiment using the Internet Movie Database dataset from the study of Pang and Lee [127], that is, unigrams, bigrams, frequent subsequences, and lemmatized frequent subtrees. For POS tagging before the step in frequent subsequence generation that splits sentences into clauses, our implementation uses the Stanford parser [128]. We use the dependency parser by Chen and Manning [122] to generate dependency trees for frequent subtree generation. For the SVM, we use the implementation from Scikit-learn with a linear kernel and all other parameters given their default values from [86]. All parameters related to frequent subsequence and subtree generation are the same as described in the study by Matsumoto et al. [110].
   - DTC: As described in the Methods section.

5.3.2 Variants of Dependency Tree Classifier

We consider the following variants of our DTC text classifier:

- DTC: as described above, with sentences not matching any pattern classified as the most common class label in the training data.
- $\mathrm{DTC}_{\mathrm{RF}}$: Sentences not matching any pattern are classified by a random forests classifier trained on the training data for each class.
- $\mathrm{DTC}_{\mathrm{CNN\text{-}W}}$: Sentences not matching any pattern are classified by a CNN-W text classifier (as defined above) trained on the training data for each class.

5.3.3 Experiments

We performed experiments with the classifiers on each class of the doctor review dataset using 10-fold cross validation. To evaluate their performance, we use weighted accuracy. For a trained classifier C and dataset D of class c_i, we define this as shown below.

$$\mathrm{WA}(C, D, c_i) = \frac{\sum_{c \in c_i} \mathrm{Accuracy}_c(C, D)}{|c_i|} \qquad (3)$$

$\mathrm{Accuracy}_c(C, D)$ is the ratio of sentences in D with class label c that were classified correctly by C. As before, $|c_i|$ is 3, the number of class labels in class c_i. We use weighted accuracy in our experiments as it places more importance on less frequent class labels, whereas regular accuracy is often above 90% because of the high number of instances labeled c_i^0 for each c_i.

The results of our experiments are shown below. In Table 54, we see that $\mathrm{DTC}_{\mathrm{CNN\text{-}W}}$ has better weighted accuracy than at least 4 baselines in each class. On average, it performs 2.19% better than the second-best method, the Matsumoto classifier ($\frac{57.05\%-55.83\%}{55.83\%}$ = 2.19%). We also observe that both the deep learning classifiers (CNN, CNN-W, and D2V-NN) and NLP classifiers (Matsumoto and DTC variants) tend to perform better than the bag-of-words classifiers (RF and SVM). This is expected as the deep learning and NLP classifiers take advantage of information in sentences such as word order and syntactic structure that cannot be expressed by a bag-of-words vector.

Table 54. Weighted accuracy of classifiers on doctor review dataset[16].

Classifier	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	**Average**
CNN	42.06%	56.69%	42.75%	51.45%	47.81%	61.42%	55.38%	60.93%	52.31%
CNN-W	49.89%	*59.68%*	44.30%	53.53%	49.71%	64.04%	54.29%	63.51%	54.87%
D2V-NN	38.83%	45.16%	38.00%	42.25%	41.44%	42.19%	41.04%	43.64%	41.57%
Matsumoto	45.76%	59.63%	45.89%	53.40%	49.89%	*66.45%*	57.24%	*68.36%*	55.83%
RF	40.78%	42.00%	34.76%	37.29%	41.62%	52.88%	45.65%	51.66%	43.33%
SVM	33.33%	35.77%	33.33%	33.33%	33.33%	48.94%	33.33%	48.07%	37.43%
DTC	51.72%	50.48%	41.27%	47.23%	38.49%	54.31%	*60.90%*	65.91%	51.29%
DTC$_{\text{RF}}$	*54.00%*	46.64%	39.19%	47.29%	40.20%	56.15%	60.57%	58.05%	50.26%
DTC$_{\text{CNN-W}}$	53.89%	59.37%	*48.66%*	*57.98%*	*50.77%*	61.43%	56.63%	67.67%	*57.05%*

Next, we further examine the performance of the top 3 classifiers, CNN-W, Matsumoto, and DTC$_{\text{CNN-W}}$. Table 55 shows the ratio of review sentences with class label c_i^x or c_i^y that were classified correctly in our experiments. Note that this is the $\text{Accuracy}_c(C, D)$ measure described above. DTC$_{\text{CNN-W}}$ generally outperforms the other classifiers with this measure; notable exceptions are c_6^y (bad staff), c_7^x (long visit time), and c_8^y (long wait time), where substantial numbers of sentences with these class labels were misclassified with the opposite label: 26.98% of c_6^y sentences were misclassified as

[16] The highest value for each c_i is italicized for emphasis.

c_6^x (good staff), 38.03% of c_7^x sentences were misclassified as c_7^y (short visit time), and 43.22% of c_8^y sentences were misclassified as c_8^x (short wait time). Finally, Table 56 shows the ratio of review sentences classified as c_i^x or c_i^y (i.e. a classifier predicted their class labels as c_i^x or c_i^y) that were classified correctly. By this measure, DTC$_{\text{CNN-W}}$ performs poorly compared with CNN-W and Matsumoto. Although the DTC algorithm's semgrex patterns classify more sentences as c_i^x or c_i^y, many of these classifications are incorrect. In the next section, we discuss reasons for some of these misclassifications.

Table 55. Per-label accuracy of top 3 classifiers on doctor review dataset for each c_i^x and c_i^y[17].

Label and classifier	$\boldsymbol{c_1}$	$\boldsymbol{c_2}$	$\boldsymbol{c_3}$	$\boldsymbol{c_4}$	$\boldsymbol{c_5}$	$\boldsymbol{c_6}$	$\boldsymbol{c_7}$	$\boldsymbol{c_8}$
$\boldsymbol{c_i^x}$								
CNN-W	31.37%	57.22%	0.00%	47.62%	40.54%	60.69%	45.07%	40.85%
Matsumoto	13.73%	57.04%	*4.00%*	48.57%	41.16%	59.16%	*52.11%*	47.89%
DTC$_{\text{CNN-W}}$	*33.33%*	*59.69%*	*4.00%*	*51.11%*	*48.02%*	*64.89%*	39.44%	*71.83%*
$\boldsymbol{c_i^y}$								
CNN-W	19.05%	27.35%	34.48%	15.44%	13.36%	35.42%	18.99%	50.75%
Matsumoto	23.81%	27.65%	35.00%	13.24%	12.93%	*43.32%*	20.25%	*57.79%*
DTC$_{\text{CNN-W}}$	*33.33%*	*48.24%*	*47.51%*	*38.97%*	*25.00%*	27.52%	*35.44%*	35.68%

Table 56. Ratio of sentences classified by the top 3 classifiers as c_i^x or c_i^y that were classified correctly.

Label and classifier	$\boldsymbol{c_1}$	$\boldsymbol{c_2}$	$\boldsymbol{c_3}$	$\boldsymbol{c_4}$	$\boldsymbol{c_5}$	$\boldsymbol{c_6}$	$\boldsymbol{c_7}$	$\boldsymbol{c_8}$
$\boldsymbol{c_i^x}$								
CNN-W	34.78%	*60.19%*	0.00%	62.50%	50.26%	66.81%	57.14%	65.91%
Matsumoto	*46.67%*	43.40%	*50.00%*	*66.23%*	*55.31%*	*71.10%*	*67.27%*	*77.27%*
DTC$_{\text{CNN-W}}$	16.04%	41.66%	10.00%	20.69%	22.58%	43.59%	23.73%	21.52%
$\boldsymbol{c_i^y}$								
CNN-W	40.00%	*41.52%*	50.56%	22.83%	*28.70%*	41.27%	29.41%	59.06%
Matsumoto	*58.82%*	34.18%	*56.52%*	*34.62%*	25.64%	*49.53%*	*53.33%*	*70.99%*
DTC$_{\text{CNN-W}}$	10.98%	13.50%	28.57%	13.38%	14.25%	22.90%	14.29%	29.96%

[17] For each c_i in this table and the next, the highest values for both c_i^x and c_i^y are italicized for emphasis.

5.4 Discussion

5.4.1 Anecdotal Examples

In this section, we show some specific patterns generated by our algorithm along with some actual review sentences that match these patterns. The semgrex pattern *{}>neg{}>>({word:/wait.*/}>{word:/long.*/})* was generated from a sentence with class label c_8^x (short wait) in class c_8 (wait time) in the doctor review dataset. It consists of a node that has 2 descendants: another generic node in a direct negation relation and *wait* in an indirect relation. The word *wait* has 1 direct descendant, the word *long*. The following is an example of a correctly matched sentence: "You are known by name and never have to wait long." This is an incorrectly matched one: "As a patient, I was not permitted to complain to the doctor about the long wait, placed on hold and never coming back to answer call." We see that it contains the words *long* and *wait*, as well as a negation (the word *never*); however, the negation is not semantically related to the long wait the author mentioned. Providing additional training data to the classifier may prevent such misclassifications by finding a pattern (or improving the rank of an existing pattern) that more appropriately makes such distinctions.

5.4.2 Limitations

In addition to the incorrect handling of negation described above, another limitation of our algorithm is that some sentences of a particular class can be sufficiently similar to sentences from another class, which may lead to misclassifications. Some examples of this can be seen in class c_6 (staff). Specifically, some sentences referring to a doctor (rather than staff members) were incorrectly classified as c_6^x (good staff) or c_6^y (bad staff). For example, “Dr. Fang provides the very best medical care available anywhere in the profession” and “Dr. Overlock treated me with the utmost respect,” which clearly refer to doctors rather than staff and should have been classified as c_6^0 (no mention of staff). The DTC algorithm generated some patterns for c_6^x that focus on positive statements for a person but miss the requirement that this person is staff. In the case of the above sentences, they were matched by *{}>>{word:/dr.*/}>>{word:/best.*/}* and *{}>>{word:/with.*/}>>{word:/dr.*/}*, respectively, which both erroneously include the word *dr*. More work is needed to address such tricky issues.

5.4.3 Conclusions

In this paper, we study the doctor review classification problem. We evaluate several existing classifiers and 1 new classifier. A key challenge of the problem is that features may be complex entities, for which polarity is not necessarily compatible with traditional positive or negative sentiment. Our proposed classifier, DTC, uses dependency trees generated from review sentences and automatically generates patterns that are then used to classify new reviews. In our experiments on a real-world doctor review dataset, we found that DTC outperforms other text classification methods. Future work may build upon the DTC classifier by also incorporating other NLP structures, such as discourse trees [129], to better capture the semantics of the reviews.

Chapter 6

Effective Social Post Classifiers on Top of Search Interfaces

Applying text classification to find social media posts relevant to a topic of interest is the focus of a substantial amount of research. A key challenge is how to select a good training set of posts to label. This problem has traditionally been solved using active learning. However, this assumes access to all posts of the collection, which is not realistic in many cases, as social networks impose constraints on the number of posts that can be retrieved through their search APIs. To address this problem, which we refer to as the *training post retrieval problem over constrained search interfaces*, we propose several keyword selection algorithms that, given a topic, generate an effective set of keyword queries to submit to the search API. The returned posts are labeled and used as a training dataset to train post classifiers. Our experiments compare our proposed keyword selection algorithms to several baselines across various topics from three sources. The results show that the proposed methods generate superior training sets, which is measured by the balanced accuracy of the trained classifiers.

6.1 Introduction

Text classification in social media is an area of active research. Examples of its application include analyzing the demographics of health-related discussions [130], inferring event attendance from non-geotagged posts [131], and detecting posts promoting extremist ideologies [132]. Training a text classification model requires a large and high-quality training set of labeled posts, where by "high quality" we generally mean that the posts should have a good coverage of the various classes, and even coverage of the various post variants within a class.

Labeling posts is generally the hardest and most expensive step in text classification applications. There has been much work on active learning, which studies how to select a good set of posts to label to achieve high training set quality. Active learning techniques assume that we have access to all documents (posts) in a collection, and iteratively select some of them to label next. However, this is not a realistic assumption in many applications, where posts' access is conducted via a constrained application programming interface (API).

In this paper, we study the *training post retrieval problem over constrained search interfaces*, which attempts to generate a high-quality posts training set, given a user-defined topic and a labeling budget. The topic is described by a few keywords that the user provides, for example for the topic "suicide" the user may provide keywords "suicide," "depressed," and "kill."

As an example application of the training post retrieval problem, consider trying to create a personal classifier for each user to filter social media posts. A user could provide a few initial keywords of interest and then the method would use these keywords to return posts for labeling. This labeling can be implicit, e.g. via clickthrough. Labeled posts can then be used to train a classifier and use it to filter further posts. Note that the initial set of keywords provided by the user are just a rough description of their latent interest profile, that is, we cannot just assume that every post that contains these keywords is relevant or that posts without these keywords are irrelevant.

The training post retrieval problem presents several challenges. The first challenge is the constrained search interface. In contrast to active learning, a method that addresses this problem does not have access to all of the available data. Instead, it must make API search queries that retrieve a limited number of posts, thus selecting the keywords that retrieve the most useful results is of key importance. Another challenge is that if we use keyword queries to generate our training set we incur coverage bias as we only get positive and negative (for binary classifiers) examples that match these queries. This problem is generally not present with active learning, where a post is picked based on how hard it is for the current classifier to classify it and not based on keywords. A third challenge is that the user-provided keywords are not perfect; there is no guarantee that any keyword provided will give 100% relevant results when used to query an API. For example, the keyword "vote" provided by a user interested in US politics may retrieve posts relevant to US politics, but may also retrieve posts relevant to voting in another country, voting on posts and/or comments on Reddit, etc.

A successful *keyword selection algorithm (KSA)* must overcome these challenges. We propose several KSAs, with the most effective being the *Top Positives Random Negatives Keyword Selection Algorithm (TPRN-KSA)*, which progressively creates keyword queries to retrieve posts, which are labelled and added to the training set. TPRN-KSA tries to achieve two goals: (a) *balance*, i.e. the number of positives and the number of negatives in the training dataset should be as close to equal as possible; and (b) *diversity*, meaning it should cover a wide range of posts within each class to properly model the data with respect to the topic of interest. We show how an algorithm has to have diversity in both the positive and the negative posts to achieve good performance.

In summary, TPRN-KSA has the following steps: First, for each input keyword it retrieves a small set of posts, which are labeled to estimate the percentage of positives for the keyword. Based on these estimations, a second portion of the budget is spent to retrieve and label more posts from the most promising (higher rate of positives) of these keywords. Finally, to address the problem of bias especially in the negative class, TPRN-KSA spends a third portion of the budget to retrieve and label a set of random posts from the API, which replace some of the biased negative posts that were retrieved during the first two steps. We carefully compute the budget for each step of the algorithm and for each keyword query to achieve the goals of balance and diversity.

A key finding in this work is that achieving diversity of the negative posts in the training set is more important that the diversity of the positive posts. Another finding is that it is not enough to add some random negative posts to the training set, but we have to also remove many or all of the biased negative posts. This may sound a little counterintuitive as more training data should be better than less. Detailed experimental evaluation shows that training a text classifier with the training set generated by TPRN-KSA outperforms state-of-the-art baseline keyword selection methods. The summary of our contributions is as follows:

- We formulate the training post retrieval problem over constrained search interfaces.

- We propose a suite of principled keyword selection algorithms, including TPRN-KSA, to solve the training post retrieval problem.
- We perform comprehensive experiments on three real datasets, which show that our proposed algorithms outperform existing baselines.
- We study the underlying reasons why the training set generated by our methods is of higher quality than the baselines. We measure the training set's balance, and the diversity of the labeled posts in both the positive and negative classes. We show how these quantities affect the quality of the training set, that is, how they are correlated to the classifier's performance.

The remainder of the paper is organized as follows: Section 6.2 discusses prior work. Section 6.3 defines the training post retrieval problem over constrained search interfaces. Section 6.4 introduces two baseline KSAs and discusses their limitations. Section 6.5 describes our initial TP-KSA method, which addresses one of the shortcomings of the baseline KSAs. Section 6.6 presents TPRN-KSA, which further refines TP-KSA to also achieve diversity among the negative posts and also balance between positives and negatives. Section 6.7 explains the experimental evaluation of our proposed method and presents the results of our experiments. We conclude in Section 6.8.

6.2 Related Work

Applying information retrieval techniques to analyze social media posts has been employed in several applications. Shen et al. developed a method to retrieve disaster event data from Twitter and other social media platforms based on event-specific hashtags [133]. Balsamo et al. proposed an information retrieval algorithm to mine data from users on Reddit by identifying subreddits relevant to opioid abuse [134]. Rao et al. proposed a neural network model specifically designed for ranking short social media posts, e.g. tweets [135]. Our proposed method differs from these in two ways. First, its goal is to retrieve a dataset of both positive and negative posts for training a text classifier rather than only posts relevant to some topic. However, the pipeline consisting of our method and a classifier could be considered an information retrieval framework in itself. Second, our proposed method is task-agnostic, i.e. it is not specifically made for any one topic or platform. We show in our experiments that our method works well across several topics and sources.

Previous work has examined collecting text data from a constrained interface using a classifier. Ruiz et al. studied how to maximize the number of relevant retrieved items using a rule-based classifier [136]. Li et al. (2013) proposed a data platform to continuously monitor the Twitter streaming API for tweets relevant to some topic using a classifier trained to detect such tweets [137]. These works are complementary to ours, as their frameworks are built around a trained text classifier which is then leveraged to gather relevant documents, whereas our work studies how to build such a classifier.

Several papers have used sets of "seed" words, phrases, or documents to find matching documents. Li et al. (2016) used seed words related to some topic and a dataset of unlabeled documents to perform dataless text classification [138]. Wang et al. proposed a technique to identify more relevant search keywords starting from an initial set of keywords for retrieving social media posts related to some topic [139]. Sadri et al. proposed a system that adapts to changes in a topic on Twitter over time by iteratively selecting phrases to track [140]. Proskurnia et al. developed a framework to extract patterns from reference documents to identify microposts related to a specific topic [141]. Li et al. (2018) developed a model for estimating the relevance between a document and a set of seed words relevant to a category using pre-trained word embeddings [142]. A limitation of relying on query keywords is that they can be a poor representation of information need [143]; we show in our experiments that our proposed method has better performance than other keyword-based methods.

Pool-based active learning uses a classifier to iteratively determine which samples from among a large dataset, e.g. a corpus of documents, are the most informative and asks a human labeler to assign a class label to them. Goudjil et al. proposed an active learning method for text classification that uses a set of SVM classifiers to determine the average posterior probability of each document within subsets of the unlabeled data [144]. Zhang et al. (2017) argued that active learning with a convolutional neural network (CNN) text classifier should focus on documents that have the most effect on the word embedding space in contrast to traditional methods such as classifier uncertainty [145]. Pool-based active learning is inapplicable to the problem we address in this paper, as the constrained search interface prevents the full dataset from being evaluated for informativeness.

Stream-based active learning involves evaluating data points (e.g. social media posts) one at a time and deciding whether to use a classifier or a human labeler to assign a class label to each one. Smailovic et al. used an SVM classifier initially trained on a Twitter sentiment dataset to perform active learning on financial-related tweets [146]. Pohl et al.

proposed a stream-based active learning method to train a classifier to detect social media posts related to crises while limiting the number of queries to human labelers [147]. Zhang et al. (2018) proposed a method to address the issue of imbalanced data in determining whether to query the human labeler by exploiting samples' second-order information [148]. This work is notably similar to ours, but one shortcoming of this method, as well as stream-based active learning methods in general, is that it relies on streaming data, e.g. Twitter’s streaming API, and thus has little control over the number of relevant posts being evaluated during the active learning process.

Positive-unlabeled (PU) learning trains a binary classifier with a dataset consisting of positive-labeled samples and additional unlabeled samples, which may be positive or negative. Li and Liu applied PU learning to text classification by combining the Rocchio method with an SVM [149]. Li et al. (2014) used PU learning to identify fake reviews on the Chinese business review website Dianping [150]. PU learning is complementary to our work, as a dataset generated by a KSA can be supplemented with additional unlabeled posts to train a classifier with PU learning instead of using traditional machine learning. However, we also note that many PU learning methods rely on the assumption that the positives are selected randomly [151], which is not applicable in this scenario.

6.3 Problem Definition

In this section, we define the *training post retrieval* problem. Given the following inputs

- List of keywords K relevant to a latent topic t
- Supervisor S, a human who labels each post as relevant or not relevant to t

- Labeling budget m, the number of posts that we submit to S for labeling

a *Keyword Selection Algorithm (KSA)* selects n pairs of (keyword query q_i, number of results r_i), $P = \{(q_1, r_1), \cdots, (q_n, r_n)\}$, to submit to the social media API, where $r_1 + \cdots + r_n = m$, as shown in Figure 21. Let $s_1, s_2, \cdots, s_n$ be the sets of posts returned by the n queries, respectively. The obtained training set is $T = s_1 \cup \cdots \cup s_n$, which is labeled by S and then used to train classifier c. The goal is to pick the P that maximizes the performance of c. *Picking the best classification method is outside the scope of this paper; we use standard SVM and CNN [84] text classifiers in our experiments.*

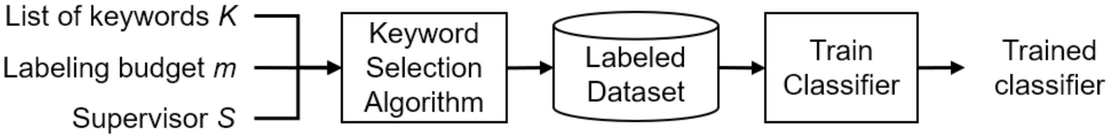


Figure 21. Process of addressing the training post retrieval problem over constrained search interfaces.

Note that n is not an input; rather it is dictated by the KSA, e.g. the algorithm may use each keyword in K to perform a query or may derive another list of keywords K' from K (e.g. a subset of K) that may be used to perform queries. P may also contain a special keyword query $q_i = \emptyset$ to mean that we obtain a random sample of posts. This does not come from a keyword in K, but is instead used by some of our methods. Support exists for retrieving random posts in a real-world setting, e.g. via the Reddit API's /random endpoint, but non-random sources of posts such as the Reddit API's /new endpoint may also suffice. Functionality similar to /new may be the only option for APIs on other social media. In general, the labeling budget m is not a constraint on the number of posts retrieved from an API, but rather the number of posts presented to the human labeler, as retrieving a large number of posts is trivial for most APIs. However, many of the KSAs evaluated in this study, including our proposed methods, retrieve only m posts.

Example: Consider a simple KSA that retrieves $\frac{m}{|K|}$ posts for each keyword in K. A user interested in posts discussing suicide provides $K =$ {"suicide", "depressed", "kill"} and $m = 600$ to the KSA, which selects $P =$ {("suicide", 200), ("depressed", 200), ("kill", 200)}. Then, for each $(q_i, r_i) \in P$, the KSA retrieves a set of posts s_i where $|s_i| = r_i$ and each post in s_i contains q_i. Next, the KSA presents the final set of posts $T = s_1 \cup s_2 \cup s_3$ to the user for labeling. The KSA then returns the labeled dataset, which is used to train a text classifier.

6.4 Baseline Keyword Selection Algorithms

In this section, we describe two simple baseline KSAs for comparison to our proposed method.

All-Keywords KSA. As described in the example above, this KSA uses every keyword in K along with a labeling budget m and supervisor S. For each keyword k, it retrieves $\frac{m}{|K|}$ posts that contain k from a keyword search API and queries S for their class labels. After all of the keywords' posts have been retrieved and labeled, it returns the labeled dataset. We expect the diversity of the negatives in this dataset to be low because every negative contains a keyword in K.

50-50 KSA. This KSA attempts to add diversity in the form of randomly retrieved posts. Like All-Keywords, it uses every keyword in K. However, instead of retrieving $\frac{m}{|K|}$ posts for each keyword k, it retrieves $\frac{m}{2|K|}$ posts that contain k for each keyword k and $\frac{m}{2}$

random posts. For example, given $K = \{$“suicide”, “depressed”, “kill”$\}$ and $m = 600$, 50-50 selects $P = \{($“suicide”, 100), (“depressed”, 100), (“kill”, 100), $(\emptyset, 300)\}$ (recall that the special keyword query $\emptyset$ retrieves random posts from an API).

Limitations: As previously mentioned, All-Keywords lacks diversity in its negatives. 50-50 attempts to address this by adding random posts, but both methods are highly dependent on the relevance of the keywords in K to achieve good balance. With All-Keywords, half of the posts retrieved by using keywords to query the API must be relevant, but with 50-50, all (or nearly all) of them should be relevant. As it is difficult to guarantee the relevance of a keyword, these KSAs often generate datasets that have poor balance.

6.5 KSA with Diversity on the Positive Samples

In this section, we describe our initial Top Positives Keyword Selection Algorithm (TP-KSA). The goal of this method is to address the balance issue due to dependence on keyword relevance experienced in All-Keywords and 50-50. Specifically, TP-KSA selects the most descriptive keywords in K, that is, keywords whose posts have a relatively high ratio of positives. It then splits its budget equally across these keywords. We also present a variant, TPP-KSA, which retrieves posts proportionally to the rate of positives for each keyword. Table 57 summarizes the notation used by our methods.

Table 57. Notation used by our methods.

Notation	Description
K	Initial list of keywords.
m	Labeling budget, i.e. the number of posts to present to S for labeling.
S	Supervisor; a human labeler to whom retrieved posts are presented for labeling.
X	List of posts to which retrieved posts are added.
y	List of class labels, where y_i is the class label for X_i.
p	List of percent positive, where p_i is the percentage of positive posts for keyword K_i.
K'	List of keywords from K retained after determining which ones to remove.
p'	List of percent positive, where p_i' is the percentage of positive posts for keyword K_i'.
b	Budget allocated equally to each keyword in K' (Equation 4).
X_k	List of posts retrieved by querying API with keyword k.
y_k	List of class labels, where y_{ki} is the class label for post X_{ki} as determined by S.
s	Number of sample posts to retrieve for each keyword in K (Equation 5).
p_k	The percentage of positive posts for keyword k.
b_i	Budget allocated proportionally to keyword K_i' (Equation 6).

6.5.1 Top Positives KSA (TP-KSA)

This method determines which keywords in K to retain according to their relevance as determined by the ratio of positive posts that they retrieve. The algorithm first expends some of its budget m to call the SampleKeyWordPosts subroutine to retrieve and label a small sample of posts from each keyword in K. It then calls the SelectKeywords subroutine to determine which keywords to retain. Next, the algorithm evenly distributes the remainder of labeling budget m, which is m minus the total number of sample posts retrieved $|X|$, among the retained keywords in K'. The budget b allocated to each remaining keyword is defined in Equation 4.

$$b = \left\lfloor \frac{m-|X|}{|K'|} \right\rfloor \qquad (4)$$

Then, for each keyword k in K', retrieve b posts that contain k from a keyword search API, ask supervisor S to label each post retrieved, and add the posts and labels to the final dataset. When this process is complete, the algorithm returns the final labeled dataset. The complete TP-KSA method is described in Algorithm 3.

Algorithm 3.

```
TP-KSA(list of keywords $K$, labeling budget $m$, supervisor $S$)
$X, y, p$ := SampleKeywordPosts($K, m, S$)
$K', p'$ := SelectKeywords($K, p$)
$b = \left\lfloor \frac{m-|X|}{|K'|} \right\rfloor$
for all keywords $k$ in $K'$ do
    $X_k$ := $b$ posts returned by querying API with keyword $k$
    $y_k$ := labels from $S$ corresponding to posts in $X_k$
    Add each post in $X_k$ to $X$
    Add each label in $y_k$ to $y$
end for
return $X, y$
```

Initial sampling: Keywords given to our method are first evaluated for relevance. For each keyword k in a list of keywords K, we query a keyword search API for a sample of posts that contain k but not contain any of the keywords previously sampled. We note that this creates the possibility that the sample posts for a keyword K_i may also contain one or more keywords K_j, where $j > i$. However, we do not exclude these posts from our sampling as this would decrease the likelihood of retrieving a positive (intuitively, posts with more than one keyword are more likely to be positive). The number of posts s in the sample is determined by the labeling budget m as shown in Equation 5.

$$s = \max\left(30, \left\lfloor \frac{m}{5|K|} \right\rfloor\right) \qquad (5)$$

For $m > 150|K|$, this formula allocates a budget of $\left\lfloor \frac{m}{5} \right\rfloor$ for all samples, which is distributed evenly between all keywords in K. For $m \leq 150|K|$, we set a minimum sample size of 30, which is considered the minimum sample size in statistics as a “rule of thumb.” In the unlikely case where $m < 30|K|$, the sample size would exceed m, so we only sample keywords until the total number of posts sampled is $0.8m$. This threshold guarantees that the entire budget m will not be spent on sampling. These values (the total sampling budget $\left\lfloor \frac{m}{5} \right\rfloor$, the minimum sample size 30, and the maximum sampling size $0.8m$) were arbitrarily selected, but our experiments show that using our proposed method with these values achieves good results for this problem. We then use supervisor S to determine whether each post in the sample is positive (relevant) or negative (not relevant). With these labels, we determine the percentage of positives for each keyword. The posts in each sample and their corresponding labels are added to the final dataset. The process of sampling keywords and determining their relevance is shown by Algorithm 4.

Algorithm 4.

```
SampleKeywordPosts(list of keywords $K$, labeling budget $m$, supervisor $S$)
$s$ := max$\left(30, \left\lfloor \frac{m}{5|K|} \right\rfloor\right)$
$X, y, p$ := empty lists
**for all** keywords $k$ in $K$ **do**
    **if** $s + |X| > 0.8m$ **then**
        **break**
    **end if**
    $X_k$ := $s$ posts returned by querying API with keyword $k$, excluding posts that contain any
    previous $k$
    $y_k$ := labels from $S$ corresponding to posts in $X_k$
    $p_k$ := percentage of positive-labeled posts in $X_k$
    Add each post in $X_k$ to $X$
    Add each label in $y_k$ to $y$
    Add $p_k$ to $p$
**end for**
**return** $X, y, p$
```

Keyword selection. Using the percentage of positive posts for each keyword, we can then determine which keywords to retain. To accomplish this, we use a method inspired by the elbow method used in clustering to determine the appropriate number of clusters in a dataset. The elbow method considers some measure, e.g. the average distance between members of a cluster [152], as a function of the number of clusters k, then chooses the k corresponding to the "elbow of the curve," i.e. the point at which the curve visibly flattens with respect to the horizontal axis, with the intuition that higher values of k offer little marginal gain. Our method takes inspiration from this approach by finding the "elbow" of a curve defined by a list of keywords and their corresponding percentages of positive posts.

Given a list of keywords K and a list of percentages p, where p_i is the percentage of positive posts in a sample of posts containing keyword K_i, this method first sorts K and p according to the values of p in descending order. Next, the method plots (i, p_i) for each i and calculates the distance between each of these points and a line drawn between the first and last point, noting the index j corresponding the the point with the greatest distance under the line. The method then retains only keywords in K with a corresponding percentage of positives $p_i > p_j$, i.e. all K_i with $i < j$. In cases where $j = 1$, which indicates that all keywords except the first one and the last one are on or above the line, the method retains every keyword in K except the last one.

As an example, consider $K = \{$Keyword 1, Keyword 2, Keyword 3, Keyword 4, Keyword 5$\}$ and $p = \{0.8, 0.75, 0.4, 0.2, 0.16\}$. As shown in Figure 22 (left), we plot (i, p_i) for each K_i in K and draw the red dotted line from the point corresponding to Keyword 1 to the point corresponding to Keyword 5 (i.e. (1, 0.8) to (5, 0.16)). We then calculate the distance from each of these points to the red dotted line and find that (4, 0.2) corresponding to Keyword 4 has the greatest distance below the line (highlighted by the blue line perpendicular to the red dotted line). We thus retain only keywords before Keyword 4.

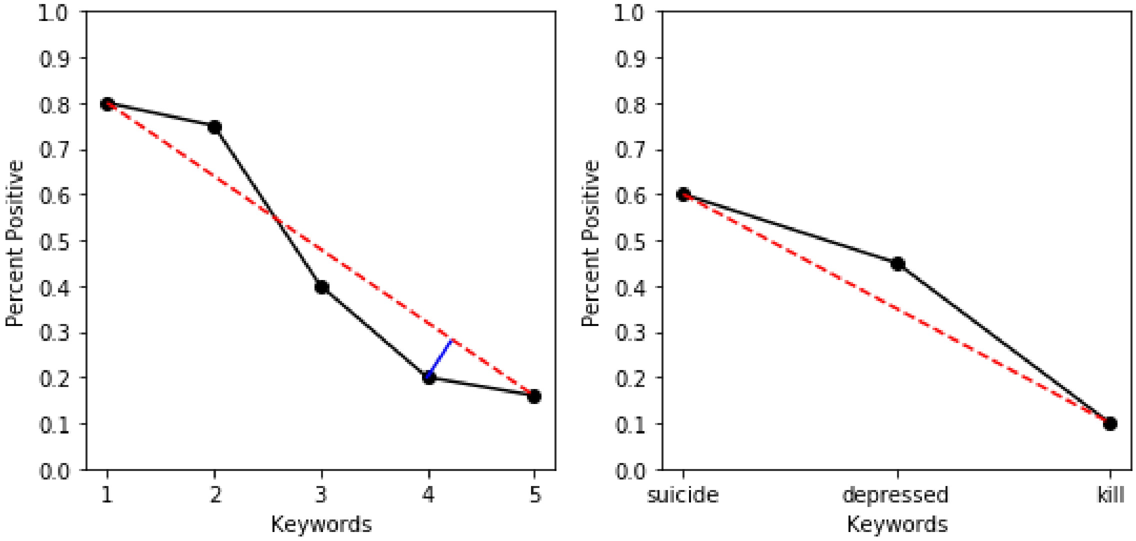


Figure 22. TP-KSA keyword selection[18].

This method of keyword selection is shown by Algorithm 5. The algorithm sorts K and p in descending order according to p, then finds the j that maximizes the distance function. The distance function calculates the distance between a point (i, p_i) and the line that intersects $(1, p_1)$ and $(|K|, p_{|K|})$ such that points below the line have a positive distance, points above the line have a negative distance, and points on the line have a distance of 0. Next, the algorithm sets $j = |K|$ if no point was below the line (signified by $j = 1$). The algorithm returns K' and p', which contain all K_i and p_i with a p_i higher than p_j. Note that the value p' (the percentages of positive posts corresponding to each keyword in K') is not used by TP-KSA, but is instead used in the variant methods described in Sections 6.5.2 and 6.6.2.

[18] Left: Keyword 4 has the greatest distance under the curve, thus the first 3 keywords are retained. Right: No keyword is under the curve, so all but the last are retained.

Algorithm 5.

```
SelectKeywords(list of keywords $K$, list of percentages $p$)
Sort $K$ and $p$ by descending order of the values in $p$
$j := \underset{i \in \{1,2,\dots,|K|\}}{\operatorname{argmax}} \frac{(p_{|K|} - p_1)(i-1) + p_i(1-|K|) + p_1(|K|-1)}{\sqrt{(p_{|K|} - p_1)^2 + (1-|K|)^2}}$
if $j = 1$ then
    $j := |K|$
end if
$K' := \{K_i \; \forall \; i \in \{1, 2, \dots, j - 1\}\}$
$p' := \{p_i \; \forall \; i \in \{1, 2, \dots, j - 1\}\}$
return $K'$, $p'$
```

Example: Given $K =$ {“suicide”, “depressed”, “kill”} and $m = 600$, TP-KSA first samples each keyword with $s = 40$ via the SampleKeywordPosts subroutine and determines that “suicide” is 60% positive, “depressed” is 45% positive, and “kill” is 10% positive (i.e. $p =$ {0.6, 0.45, 0.1}). TP-KSA then continues with the SelectKeywords subroutine and determines $j = 3$ corresponding to the keyword “kill” as shown in Figure 22 (right). Thus $K' =$ {“suicide”, “depressed”} and each keyword in K' is allocated a budget $b = 240$, so $P =$ {(“suicide”, 40), (“depressed -suicide”, 40), (“kill -suicide -depressed”}, 40), (“suicide”, 240), (“depressed”, 240)}. “k_i -k_j” denotes a keyword query for posts that contain keyword k_i but do not contain keyword k_j.

6.5.2 Top Positive Proportional KSA (TPP-KSA) Variant

We also propose a variant to TP-KSA that aims at retrieving more positive posts by allocating a proportionally higher budget to keywords with higher percentages of positive sample posts instead of allocating the same budget to each keyword in K'. More specifically, the method uses the list of percentages p' to determine the budget b_i for a keyword K'_i as shown in Equation 6.

$$b_i = \left\lfloor p'_i \frac{m - |X|}{\sum_{i=1}^{|K'|} p'_i} \right\rfloor \qquad (6)$$

For each keyword K'_i in K', we then use this new budget b_i to retrieve posts containing K'_i similarly to TP-KSA. The TPP-KSA method is shown in Algorithm 6.

Algorithm 6.

```
TPP-KSA(list of keywords $K$, labeling budget $m$, supervisor $S$)
$X, y, p$ := SampleKeywordPosts($K$, $m$, $S$)
$K'$, $p'$ := SelectKeywords($K$, $p$)
for all keywords $K'_i$ in $\$K'$ do
    $b_i := \left\lfloor p'_i \frac{m - |X|}{\sum_{i=1}^{|K'|} p'_i} \right\rfloor$
    $X_k$ := $b_i$ posts returned by querying API with keyword $K'_i$
    $y_k$ := labels from $S$ corresponding to posts in $X_k$
    Add each post in $X_k$ to $X$
    Add each label in $y_k$ to $y$
end for
return $X$, $y$
```

6.6 Extend TP-KSA to Add Diversity on the Negative Samples and Balance

While TP-KSA increases the diversity in the positive portion of its generated dataset, the posts in the negative portion each contain at least one keyword. This bias results in low diversity in that portion of the dataset. TP-KSA also does not sufficiently address the balance problem that All-Keywords and 50-50 have; the relevance of the keywords it is given remains the most significant factor in determining how well-balanced the resulting dataset is. To resolve these issues, we created TPRN-KSA, which builds upon TP-KSA by (a) discarding negative posts containing keywords to eliminate the source of bias and replacing them with randomly selected posts to add diversity to the negative samples, and (b) aiming for the same number of positives and negatives in the final dataset.

6.6.1 Random Negatives Variant of TP-KSA (TPRN-KSA)

We first assume that, for retrieval purposes, all randomly selected posts are negative. Then, to balance the positives and negatives (recall that we discard negatives returned by keyword queries), we need to compute the total number m_k of posts that should be retrieved using keywords versus the number $m - m_k$ of posts that should be retrieved randomly. Hence, we have:

$$m_k \cdot \text{average}(p') = m - m_k \qquad (7)$$

where the left side of Equation 7 represents the target number of positive posts in the final dataset, while the right side is the number of negative (random) posts.

m is our overall labeling budget, $\text{average}(p')$ is the average of all the values in p', and m_k is the budget for posts containing a keyword. Equation 8 shows the derived formula for m_k, which also incorporates a maximum value of $0.8m$ to ensure we avoid $m_k = m$, i.e. no budget remains for random posts.

$$m_k = \left\lfloor \min\left(\frac{m}{1 + \text{average}(p')}, 0.8m\right) \right\rfloor \qquad (8)$$

We now describe our TPRN-KSA method. Like TP-KSA, it first calls the SampleKeyWordPosts and SelectKeywords subroutines. Then, using the list of percentages of positive posts p' returned by SelectKeywords, we calculate m_k using Equation 8. The algorithm then retrieves and elicits labels for b posts for each of the keywords in K' as in TP-KSA, but with a value of b that incorporates m_k:

$$b = \left\lfloor \frac{m_k - |X|}{|K'|} \right\rfloor \qquad (9)$$

Next, the posts with negative labels are removed, and replaced with $m - m_k$ random posts retrieved from a keyword search API. While these random posts were assumed to be negative in our derivation of m_k, there may be positives among them in practice, thus they must be labeled by supervisor S. After these posts are labeled, they are added to the final dataset. TPRN-KSA is further described in Algorithm 7. It is also important to note that TPRN-KSA returns fewer than m labeled posts. However, we show in our experiments that despite generating dataset that is smaller compared to TP-KSA and the baselines, TPRN-KSA generates a dataset that leads to better classifier performance than those methods.

Algorithm 7.

```
TPRN-KSA(list of keywords $K$, labeling budget $m$, supervisor $S$)
$X, y, p$ := SampleKeywordPosts($K, m, S$)
$K', p'$ := SelectKeywords($K, p$)
$m_k$ := $\lfloor \min(\frac{m}{1 + \text{average}(p')}, 0.8m) \rfloor$
$b$ := $\left\lfloor \frac{m_k - |X|}{|K'|} \right\rfloor$
for all keywords $k$ in $K'$ do
    $X_k$ := $b$ posts returned by querying API with keyword $k$
    $y_k$ := labels from $S$ corresponding to posts in $X_k$
    Add each post in $X_k$ to $X$
    Add each label in $y_k$ to $y$
end for
Remove all negative labels from $y$ and remove their corresponding posts from $X$
$X_\emptyset$ := $m - m_k$ posts returned by querying API with $\emptyset$
$y_\emptyset$ := labels from $S$ corresponding to posts in $X_\emptyset$
Add each post in $X_\emptyset$ to $X$
Add each label in $y_\emptyset$ to $y$
return $X, y$
```

Example: Given $K = \{$“suicide”, “depressed”, “kill”$\}$ and $m = 600$, TPRN-KSA first determines K' and p'. Then, it calculates $m_k = 393$ and each keyword in K' is allocated $b = 136$ to retrieve additional posts with those keywords. Of all the retrieved posts (including those retrieved by SampleKeywordPosts), 180 are positive. The remaining 213 posts are removed from X and their corresponding class labels are removed from y. TPRN-KSA then retrieves $m - m_k = 207$ random posts for a total dataset size of 387 and $P = \{$(“suicide”, 40), (“depressed -suicide”, 40), (“kill -suicide -depressed”, 40), (“suicide”, 136), (“depressed”, 136), ($\emptyset$, 207)$\}$.

6.6.2 TPPRN-KSA: Add Random Negatives to TPP-KSA

Our experiments also include a variant that combines both the TPP-KSA and TPRN-KSA variants called TPPRN-KSA. As with TPRN-KSA, this method calculates a labeling budget m_k for posts with a keyword. It then proportionally allocates budgets to each keyword in K' as in TPP-KSA, but uses m_k instead of m when calculating these budgets. For brevity, we do not include the psuedocode of this variant, but it is composed of lines 1-3 of Algorithm 7, then lines 3-9 of Algorithm 6 (with m_k replacing m in line 4), followed by lines 11-16 of Algorithm 7.

6.7 Experiments

In this section, we describe our experimental evaluation of our proposed method compared to several baselines.

6.7.1 Data

We use data from three sources in our experiments. In each experiment, posts from one of 15 message boards (DailyStrength) or one of 20 subreddits (Reddit), or news headlines from one of 35 categories (The Huffington Post) are labeled positive; all other posts/headlines in our collected data from that source are labeled negative. Note that while the datasets for our experiments have been downloaded in advance, this is not a requirement of our proposed method. We use these datasets in our experiments to simulate keyword search APIs from which posts are retrieved.

DailyStrength. DailyStrength is a social networking site that enables patients to exchange experiences and treatments, discuss daily struggles and successes, and receive emotional support [153]. We collected DailyStrength posts from 2006 to 2019 and determined the top 50 message boards by number of posts collected. From these 50, we manually selected 15 message boards related to medical conditions. These 15 message boards, as well as the number of positives and negatives in their respective datasets and the keywords used in our experiments, are listed in Table 58. Each post in these datasets is the first post in a thread.

Table 58. DailyStrength datasets.

Message board	Positives	Negatives	Keywords
Alcoholism	2077	511,988	drinking sober drink alcohol aa
Anxiety	3418	510,647	anxiety anxious panic attacks attack
Chronic Fatigue Syndrome	1611	512,454	cfs fatigue chronic illness energy
Chronic Pain	1692	512,373	pain chronic knee meds norco
Depression	2881	511,184	depression feel don['t] depressed anymore
Eating Disorders	1761	512,304	eating eat binge weight ed
Gastritis	1642	512,423	gastritis stomach ppi endoscopy acid
Graves' Disease	1622	512,443	graves tsh thyroid methimazole labs
Hidradenitis Suppurativa	1660	512,405	hs groin armpots boils dermatologist
Insomnia	1616	512,449	sleep insomnia sleeping asleep night
Multiple Personalities	1669	512,396	alters alter personalities therapist parts
Myasthenia Gravis	1764	512,301	mg mestinon prednisone myasthenia neuro
Obsessive Compulsive Disorder	1634	512,431	ocd thoughts intrusive obsessions anxiety
Post-Traumatic Stress Disorder	3956	510,109	ptsd trauma tw nightmares triggered
Pulmonary Embolism	1632	512,433	pe clots warfarin clot xarelto

The Huffington Post. We downloaded the News Category Dataset [154], a collection of news headlines in 41 categories published by The Huffington Post between 2012 and 2018. We noted that some of these categories were similar to each other and combined them:

- Arts, Arts & Culture, and Culture & Arts
- Healthy Living and Wellness
- Parenting and Parents
- Style and Style & Beauty
- Worldpost and The Worldpost

Combining these categories and adding the remaining categories from [154] gave us a total of 35 as shown in Table 59. Each headline in the categories' datasets is represented by the concatenation of the headline itself and a short description of the headline's article.

Table 59. Huffington Post datasets.

Category	**Positives**	**Negatives**	**Keywords**
Arts	3878	196,975	art artist artists imageblog exhibition
Black Voices	4528	196,325	black police african racial racism
Business	5937	194,916	business company companies ceo wall
College	1144	199,709	college students campus university colleges
Comedy	5175	195,678	colbert snl jimmy maher stephen
Crime	3405	197,448	police suspect man shooting allegedly
Divorce	3426	197,427	divorce divorced ex marriage split
Education	1004	199,849	education students school teachers schools
Entertainment	16,058	184,795	movie film trailer star actor
Environment	1323	199,530	animal climate week weather animals
Fifty	1401	199,452	aging retirement age 50 older
Food & Drink	6226	194,627	recipes recipe food cooking taste
Good News	1398	199,455	dog homeless rescue trump pit
Green	2622	198,231	climate change environmental california dog
Healthy Living	24,521	176,332	health trump study life sleep
Home & Living	4195	196,658	home photos ideas diy craft
Impact	3459	197,394	homeless world homelessness people global
Latino Voices	1129	199,724	latino latinos latina puerto mexican
Media	2815	198,038	news fox media cnn journalists
Money	1707	199,146	credit money financial tax debt
Parenting	12,632	188,221	kids parents children mom child
Politics	32,739	168,114	trump donald gop clinton president
Queer Voices	6314	194,539	gay queer lgbt lgbtq trans
Religion	2556	198,297	pope francis church faith god
Science	2178	198,675	scientists nasa space science mars
Sports	4884	195,969	nfl football nba game player
Style	11,903	188,950	photos fashion style check tumblr
Taste	2096	198,757	recipes delicious food make cooking
Tech	2082	198,771	apple google iphone facebook tech
Travel	9887	190,966	travel hotels photos hotel travelers
Weddings	3651	197,202	wedding weddings marriage bride married
Weird News	2670	198,183	man cops dog fark weird
Women	3490	197,363	women funniest feminist woman sexual
World News	2177	198,676	korea north rogingya myanmar korean
Worldpost	6243	194,610	isis syria syrian minister turkey

Reddit. We collected Reddit posts from 2008 to 2018 from 72 primarily health-related subreddits. From these, we randomly selected 16 subreddits related to medical conditions with at least 1,000 posts and added four additional subreddits not specifically related to a medical condition: /r/Dentistry, /r/electronic_cigarette, /r/politics, and /r/SuicideWatch. These 20 subreddits, as well as the number of positives and negatives in their respective datasets, are listed in Table 60. Each post in these datasets is a “self post,” i.e. a post that contains text rather than a link.

Table 60. Reddit datasets.

Subreddit	**Positives**	**Negatives**	**Keywords**
/r/ADHD	75,953	1,297,505	adhd adderall vyvanse medication focus
/r/anxiety	106,253	1,267,205	anxiety panic anxious attack attacks
/r/aspergers	35,037	1,338,421	aspergers aspie autism asperger aspies
/r/Asthma	3455	1,370,003	asthma inhaler albuterol inhalers ventolin
/r/BipolarReddit	17,271	1,356,187	bipolar manic mania lamictal lithium
/r/BPD	41,647	1,331,991	bpd dbt relationship fp borderline
/r/cancer	14,999	1,358,459	cancer chemo radiation tumor oncologist
/r/cfs	6287	1,367,171	cfs fatigue chronic syndrome symptoms
/r/ChronicPain	14,698	1,358,760	pain chronic nerve doctor disc
/r/Dentistry	40,524	1,332,934	dentist teeth tooth dental wisdom
/r/Depression	202,464	1,170,994	depression feel depressed life friends
/r/diabetes	26,471	1,346,987	diabetes insulin diabetic sugar pump
/r/electronic_cigarette	112,137	1,261,321	mod vape coils coil tank
/r/Hemophilia	1737	1,371,721	hemophilia factor hemophiliac hemophiliacs bleeds
/r/Infertility	20,609	1,352,849	infertility ivf iui cycle moan
/r/kidneystones	1166	1,372,292	stone kidney stones pain ureter
/r/politics	13,281	1,360,177	2018 trump politics just html
/r/STD	11,900	1,361,558	sex herpes std penis unprotected
/r/SuicideWatch	114,484	1,258,974	life suicide kill die want
/r/thritis	1392	1,372,066	arthritis ra rheumatologist pain joints

6.7.2 Baselines

We compare our proposed method to five baselines. As described in our problem definition, these methods take a list of keywords K, a labeling budget m, and a supervisor S as input. For brevity, we use “posts” here and in the remainder of the paper to refer to both posts from DailyStrength or Reddit, and news headlines and their article summaries from the Huffington Post.

- **All-Keywords:** As described in Section 6.4.
- **50-50:** As described in Section 6.4.
- **Double Ranking:** This method uses the Double Ranking method proposed by Wang et al. [139]. The keywords in K are first used to retrieve and label sample posts with Algorithm 4. The posts are then split according to their labels into X_{pos} and X_{neg}. These two datasets, along with K and a list of stopwords defined by Gensim [125], are passed to the Double Ranking algorithm, which runs for two iterations to produce a new list of keywords K', where $|K'| = |K|$. These keywords are used to retrieve and label additional posts as in lines 3-10 of Algorithm 3. An important distinction with this method is that it retrieves much more than m posts, as the Double Ranking algorithm retrieves 300[19] posts per keyword in each iteration. However, only m posts are presented to S for labeling and added to the final dataset.
- **Active Learning:** This method uses pool-based active learning with entropy as an uncertainty measure. The initial dataset is made by iteratively querying each of the keywords in $K \cup \emptyset$ for one post one at a time until the dataset has at least one post of each label. Then, using a pool of 100,000 posts, each step of the active learning

[19] This value was used by the experiments in [139].

process presents the $\frac{m}{10}$ posts in the pool with the highest entropy to S for labeling and adds them to the dataset until the dataset has m labeled posts.

- **Random:** Retrieve m by querying an API with $\emptyset$ (i.e. retrieve m random posts) and present them to S for labeling, then return the labeled dataset.
- **Ideal:** Retrieve $\frac{m}{2}$ positive posts and $\frac{m}{2}$ negative posts from an API. As this method is capable of retrieving posts based on their label, it cannot be applied to a real-world setting. Instead, it serves as an approximate upper limit for comparison to the other methods.

6.7.3 Experiment Setup

Each experiment takes the following input:

- A KSA M
- A dataset D consisting of positives from one message board, subreddit or topic, and negatives from the remaining data from the same source
- A labeling budget m
- A number of keywords n (in all of our experiments, n = 5)

Each experiment is run as follows:

1. Remove a random sample from D of size $0.2|D|$ for use as test data.
2. Determine the top n words according to their information gain according to the data and labels in D; these words will be used as a list of keywords K.
3. Create a simulated keyword search API A using the data from D.

4. Run M with m and K as input. The supervisor S required by M will be simulated by the dataset's existing labels.
5. Train a text classifier C using the labeled data from M.
6. Use C to classify the test data (see Step 1) and record the results.

We then combine the all of the results for one source and one method and report the average balanced accuracy [88], e.g. the average balanced accuracy across all subreddit datasets in the Reddit data for TP-KSA, for $m = 100, 200, \cdots, 1000$. This metric was chosen because accuracy can be misleadingly high when applied to the results produced from our test datasets, which are highly imbalanced toward the negative class.

These experiments are performed with both a CNN text classifier and a linear SVM with TF-IDF vectorization. Our CNN classifier splits the labeled data into training (80% of the data) and validation (20% of the data) datasets, performs 50 training epochs with a batch size of 100, uses window sizes of 3, 4, and 5 with 100 filters each, and uses 300-dimension fastText [85] embeddings pre-trained on Wikipedia. As we are using pre-trained embeddings, the posts selected by each KSA in our experiments do not affect the embedding space.

Our SVM experiments use the implementation in Scikit-learn [86] with 1000 features representing n-grams of sizes 1-3 with a minimum document frequency of 3% and stop words removed. All other parameters are set to their default values in Scikit-learn. Additionally, our SVM experiments are performed with 5-fold cross-validation (i.e. D is split into 5 disjoint subsets, the experiment process described above is performed 5 times with one of the subsets used as the test data, and the results are averaged).

6.7.4 Results – Balanced Accuracy

We split our experimental evaluation into three steps. We first compared the balanced accuracy of classifiers built from training data generated by each of the four variants of our proposed method. We then compared the best of these to the baselines. Finally, using the results of these experiments, we observe the effect of balance and diversity on classifier performance and report our findings in Section 6.7.5.

Comparison of TP-KSA Variants

Figure 23 shows the average balanced accuracy of SVM and CNN classifiers trained with data from each of the four TP-KSA variants. From these results, it is clear that the two “random negatives” variants, TPRN-KSA and TPPRN-KSA, lead to a classifier with higher balanced accuracy than the other two methods. However, the balanced accuracy of each of the two “proportional” variants (TPP-KSA and TPPRN-KSA) is very close to the balanced accuracy of its non-proportional counterpart. For the remainder of our experiments, we use TPRN-KSA rather than TPPRN-KSA as the additional functionality of the latter confers no significant benefit.

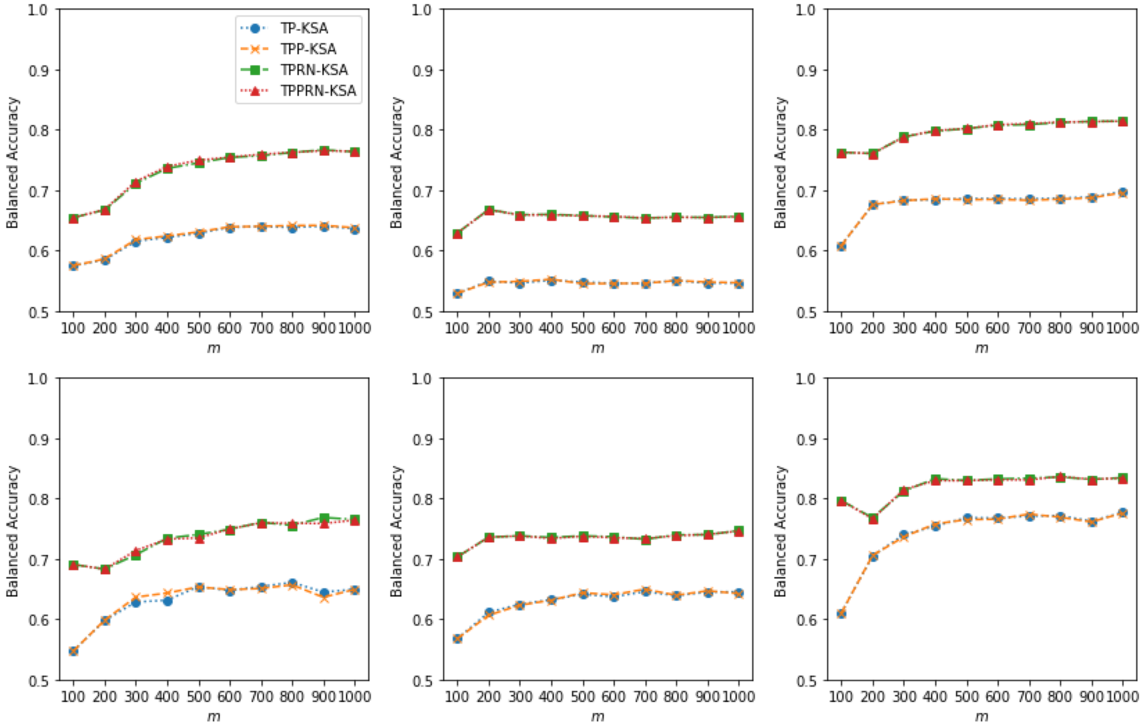


Figure 23. Balanced accuracy of each TP-KSA variant using SVM (top) and CNN (bottom)[20].

Comparison to Baselines

The results of the experiments with TPRN-KSA and the baselines are shown in Figure 24. From these results we see that TPRN-KSA has higher balanced accuracy than all real-world baselines in all three sources with both SVM and CNN. The only method with higher balanced accuracy is the Ideal baseline, which has the benefit of being able to retrieve posts by their class label. All-Keywords, 50-50, Double Ranking, and Active Learning have similar balanced accuracy. Among these four baselines, All-Keywords tends to perform the best, while Active Learning tends to perform the worst.

[20] Left: DailyStrength. Middle: The Huffington Post. Right: Reddit.

Unsurprisingly, the Random method has the lowest balanced accuracy. In the following section, we will further analyze the performance of these methods within the context of balance and diversity.

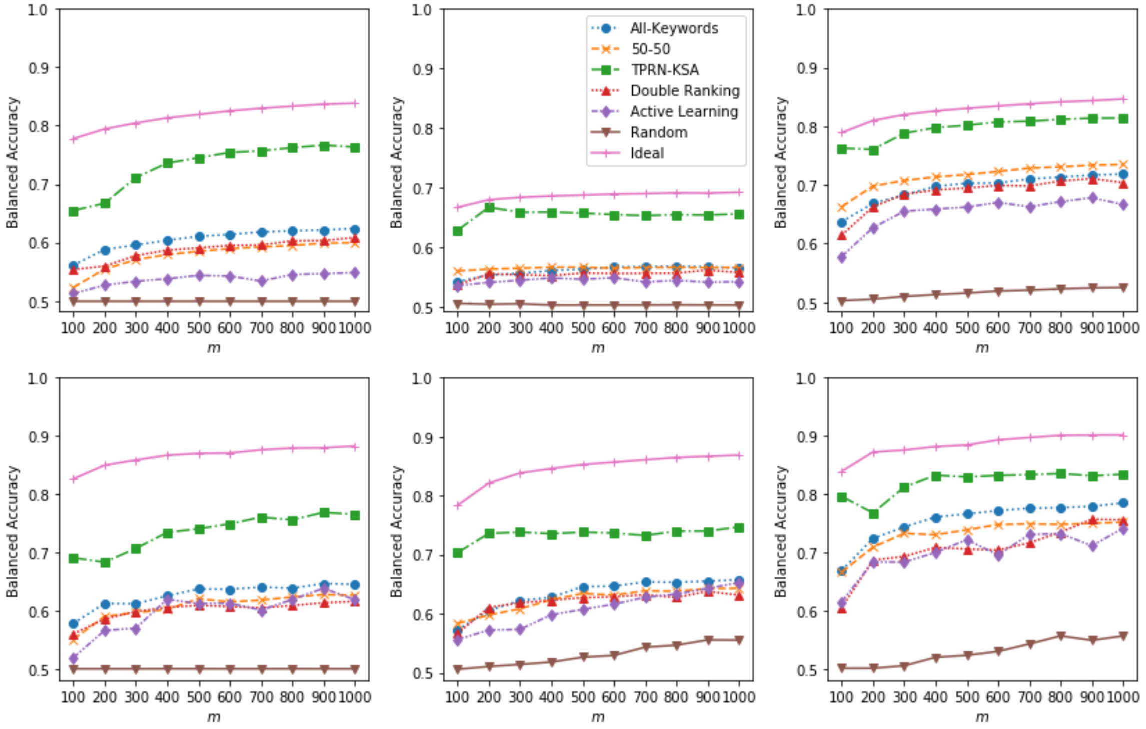


Figure 24. Balanced accuracy of TPRN-KSA vs. baselines using SVM (top) and CNN (bottom)[21].

6.7.5 Results – Effect of Balance and Diversity on Classifier Performance

Recall that we previously stated that a classifier training dataset must achieve both balance and diversity to maximize classifier performance. More precisely, given a binary classifier trained on a training dataset T, the classifier's balanced accuracy on a test dataset should positively correlate to a combination of the balance and diversity of T. To

[21] Left: DailyStrength. Middle: The Huffington Post. Right: Reddit.

determine this correlation, we must first quantify balance and diversity. To do so, we will use 3 additional metrics:

- **Percent Positive**. The ratio of positive posts in T.
- $\mathbf{KL_{pos}}$ and $\mathbf{KL_{neg}}$. The Kullback-Leibler divergence [155] of the positives (or negatives) in T from an equal number of positives (or negatives) taken from D. A bag-of-words model is used to represent each thread in these calculations. The divergence of part of T from a random sample with the same label acts as a measure of diversity in T, where lower values correspond to more diversity in that part of T.

Unless otherwise specified, our discussion of these metrics focuses on the SVM experiments, as the use of cross-validation makes these results more reliable. We report the averages of these metrics for each source in Figure 25 for SVM as determined by the results of our experiments in Section 6.7.4. Note that with the exception of Active Learning, which incorporates a classifier into its selection of posts, these results differ from the CNN results only due to the use of cross-validation in the SVM experiments.

When comparing the percent positive of the datasets generated by each method we see that TPRN-KSA is the closest to 50% positive with data from DailyStrength and The Huffington Post. With the Reddit data, the Double Ranking method is closer for some values of m and All-Keywords is above, but also close to, 50% positive. The percent positive of the Random method is very low, which is likely the most substantial contributing factor to its low balanced accuracy.

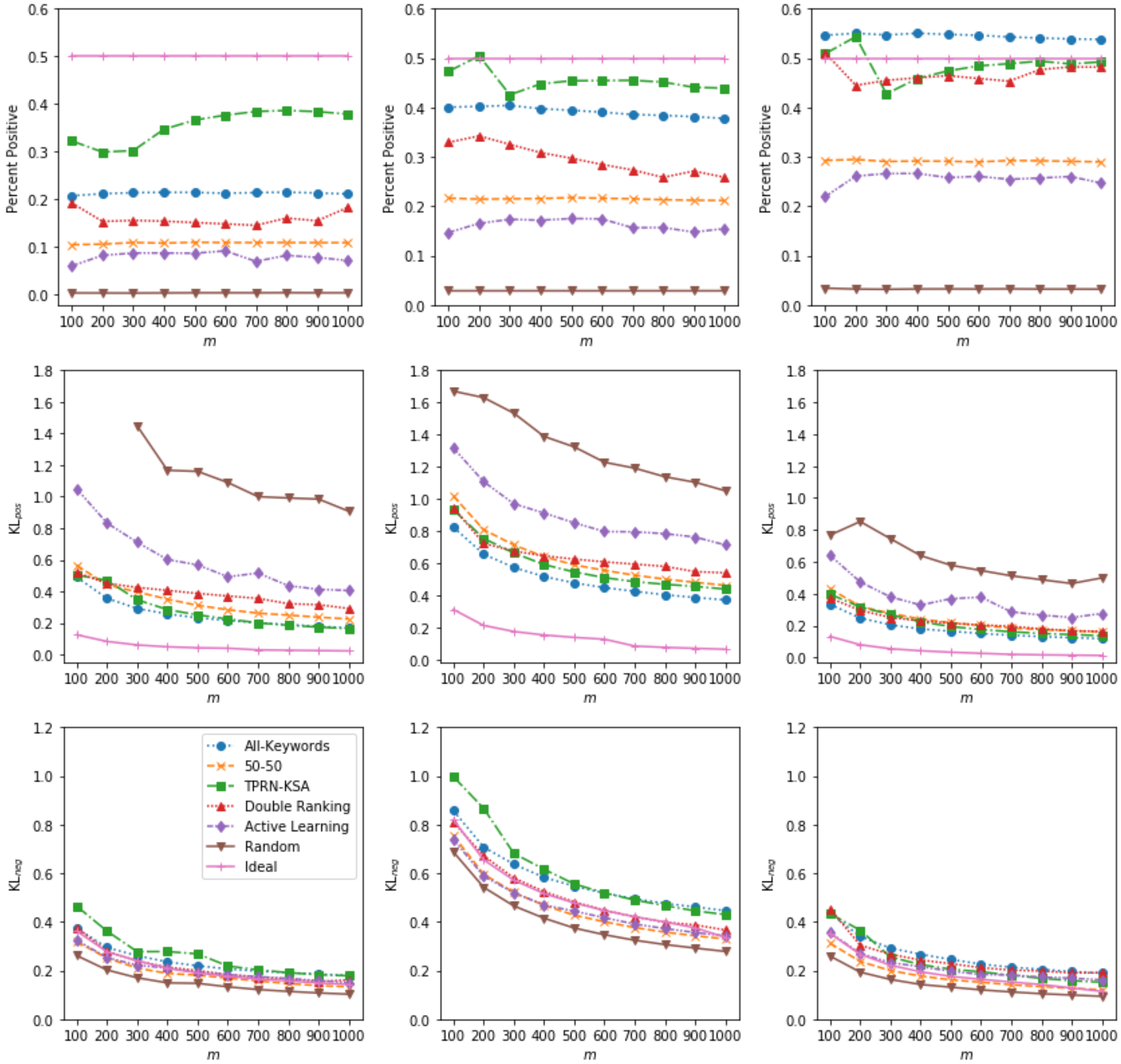


Figure 25. Percent positive, KL$_{\text{pos}}$, and KL$_{\text{neg}}$, as determined in the SVM experiments[22].

We also observed that each method's percent positive relative to other methods is generally opposite the same method's relative KL$_{\text{pos}}$, i.e. if method a has a higher percent positive than method b, then method a will tend to have a lower KL$_{\text{pos}}$ than method b.

[22] Left: DailyStrength. Middle: The Huffington Post. Right: Reddit.

Notably, two exceptions to this are TPRN-KSA and Double Ranking, which tend to have slightly higher KL_{pos} than All-Keywords and 50-50, respectively. This may be due to Double Ranking's retrieved positive posts becoming more topical (and thus less diverse) due to the refined keywords it uses, while the TP-KSA methods' use of a subset of K sacrifices some of the positive diversity provided by All-Keywords for better balance. The KL_{pos} for the Random method with the DailyStrength dataset is not given for $m < 300$, as there were an insufficient number of positives to calculate a meaningful result with these values of m.

The KL_{neg} of each generated dataset tends to be directly proportional to that dataset's percent positive because a higher number of positive posts means the dataset has fewer negatives and thus less opportunity for diversity. TPRN-KSA also follows this trend for smaller values of m, but its KL_{neg} falls below one or more other methods as m increases, particularly with the Reddit data. This suggests that the benefit conferred by TPRN-KSA's use of random negatives is more pronounced as m increases. We also observed that 50-50 tends to have slightly lower KL_{neg} than Active Learning, which may be explained by the fact that the negatives retrieved by 50-50 are random, while those retrieved by Active Learning are instead selected according to their entropy.

These metrics generally follow the same trends in the CNN results, shown in Figure 26, as they do in the SVM results. A notable exception is Active Learning, which has higher KL_{neg} than all other methods except for Random. As suggested above, the non-random nature of the negatives selected by Active Learning has an adverse effect on its KL_{neg}. This is amplified by the assertion by Zhang et al. (2017) [145] that active learning selection with a multi-layered neural model should be based on the embedding space rather than entropy. However, while the SVM and CNN results are not directly comparable since only the SVM experiments used cross-validation, we observe that the balanced accuracy of Active Learning in the CNN experiments is higher relative to other methods than it is the SVM experiments, suggesting that $\mathrm{KL}_{\mathrm{neg}}$ has less of an impact on classifier performance, at least when active learning is used.

Next, we combine these three metrics and study their correlation to classifier performance. We first normalize them and then take the harmonic mean. The intuition behind using their harmonic mean is that we want to show that a classifier tends to perform well when all three of these metrics are high, but performs worse when one or more is low. Because the harmonic mean gives more weight to smaller values, it tends to be lower when one value is low compared to the arithmetic or geometric means. The normalized balance is $B = 1 - 2|0.5 - p|$ where p is the percent positive and has range [0, 1]. The normalized diversity, which is defined separately for the positive and negative portions of a generated dataset, is $D_x = e^{-KL_x}$, where x is either “pos” or “neg,” and has range [0, 1]. The harmonic mean of these three values is then defined as follows:

$$H = \left(\frac{B^{-1} + D_{\mathrm{pos}}^{-1} + D_{\mathrm{neg}}^{-1}}{3}\right)^{-1} \qquad (10)$$

Using this definition, we calculated each experiment's H and compared it to the same experiment's classifier balanced accuracy to study their correlation.

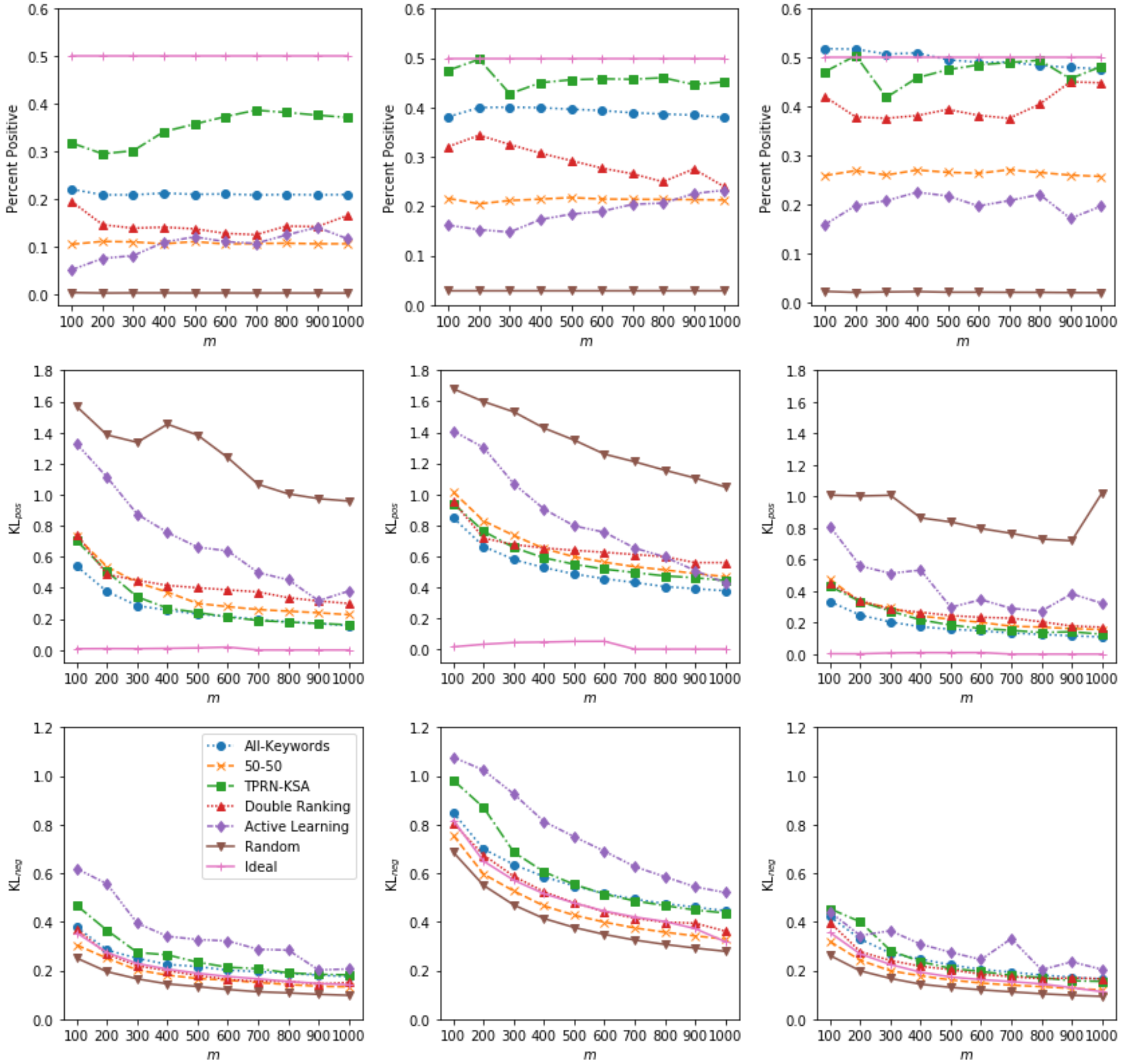


Figure 26. Percent positive, KL$_{pos}$, and KL$_{neg}$, as determined in the CNN experiments[23].

[23] Left: DailyStrength. Middle: The Huffington Post. Right: Reddit.

Balanced Accuracy vs. Balance and Diversity

We plotted the results of each of our previous experiments for each source in two dimensions. Each experiment is represented by coordinates (H, A), where H is the harmonic mean of the balance, the diversity of the positives, and the diversity of the negatives of the training dataset generated in that experiment and A is the balanced accuracy the trained classifier achieved on the test data in that experiment. Experiments where the harmonic mean is undefined are excluded from our analysis. We show the plotted experiments in Figure 27. We also show the Pearson correlation coefficient [51] of balanced accuracy and the harmonic mean of balance and diversity for each source in Table 61. All three sources show that a classifier's balanced accuracy is strongly correlated with the balance and diversity of the dataset that was used to train the classifier. However, we also note that these correlations vary substantially between sources and classifiers; this suggests that other factors beyond balance and diversity play a role in classifier performance.

Table 61. Correlation of classifier balanced accuracy and the harmonic mean of balance and diversity of classifier training data.

Source	***n***	***r* (SVM)**	***p* (SVM)**	***r* (CNN)**	***p* (CNN)**
DailyStrength	1500	0.8629	< 0.001	0.8223	< 0.001
The Huffington Post	3500	0.6610	< 0.001	0.7715	< 0.001
Reddit	2000	0.7411	< 0.001	0.6807	< 0.001

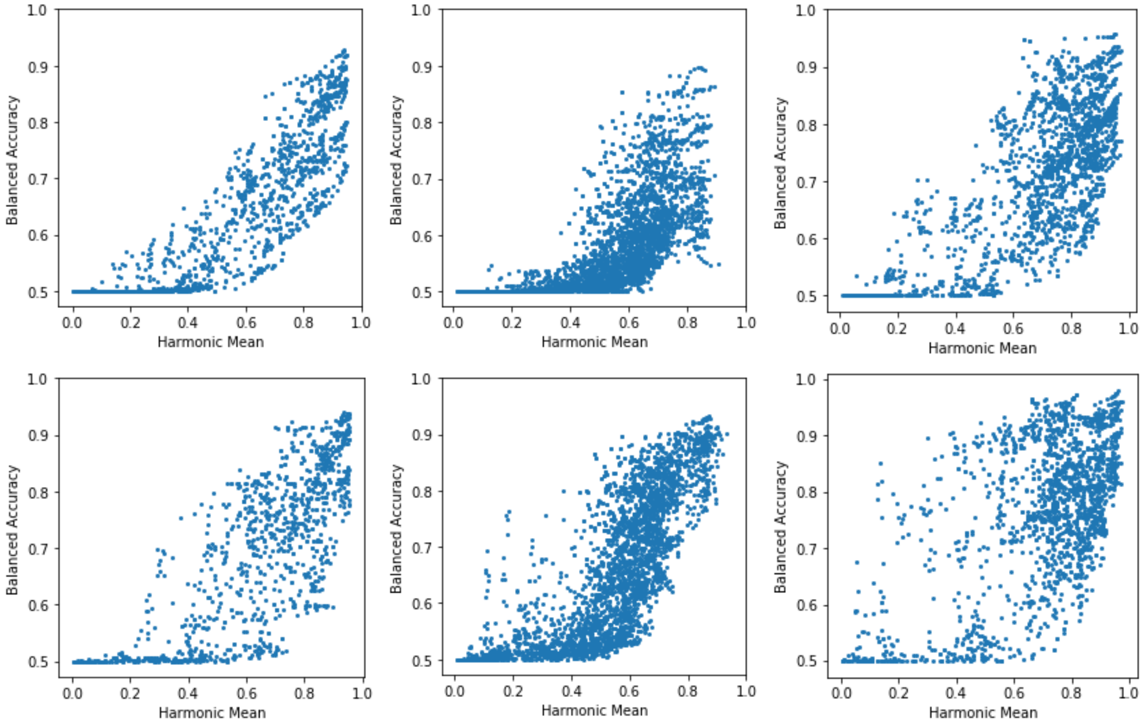


Figure 27. Correlations between balanced accuracy and the harmonic mean of balance and diversity with SVM (top) and CNN (bottom)[24].

Limitations

Despite the performance of TPRN-KSA, it has several limitations. First, its sampling of posts to determine the percentage of positives for each keyword (Algorithm 4) does not account for the fact that a keyword may appear in a post that was sampled using another keyword. Accounting for this could reduce the amount of the budget used for sampling. Another issue with the sampling of posts is its use of arbitrary values; specifically, reserving 20% of m for sampling, applying the "rule of 30" for determining minimum sample sizes, and imposing a maximum sample size of $0.8m$ for small values

[24] Left: DailyStrength. Middle: The Huffington Post. Right: Reddit.

of m. Similarly, TPRN-KSA's keyword selection (Algorithm 5) is inspired by the elbow method, which is a generally unreliable means of determining the number of clusters in a dataset. A more principled approach to these methods may lead to better results.

6.8 Conclusions

We proposed the training post retrieval problem over constrained search interfaces. We also proposed a method to address this problem, TPRN-KSA. This method is built on the assumption that balance and diversity in a training dataset positively affect the balanced accuracy of a classifier trained with the data. TPRN-KSA outperformed all other variant methods and several baselines in our experiments. For $m = 1000$, TPRN-KSA has an improvement of 13.96%, 8.95%, and 7.91% with SVM and 11.90%, 8.94%, and 4.92% with CNN over the best baseline for DailyStrength, The Huffington Post, and Reddit, respectively. We followed up these experiments with an analysis on the correlation between classifier balanced accuracy and the harmonic mean of the balance and diversity of training data. We found that they were positively correlated, supporting our initial assumption. Future work may address the limitations in TPRN-KSA as discussed above by improving the sampling behavior and proposing new methods for keyword selection, e.g. by incorporating other metrics.

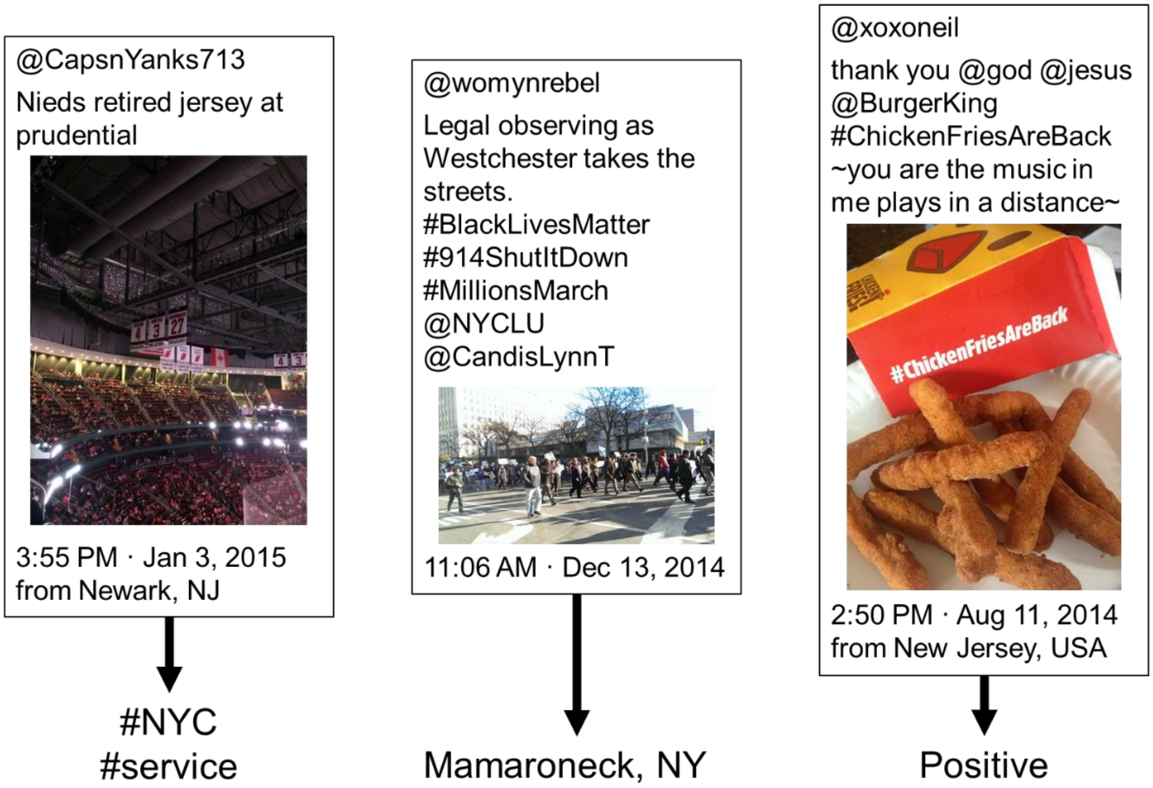
Chapter 7

Holistic Embedding Generation for Twitter Machine Learning Applications

Twitter is a frequent target for machine learning research and applications. Many problems, such as sentiment analysis, image tagging, and location prediction have been studied on Twitter data. Much of the prior work that addresses these problems within the context of Twitter focuses on a subset of the types of data available, e.g. only text, or text and image. However, a tweet can have several additional components, such as the location and the author, that can also provide useful information for machine learning tasks. In this work, we explore the problem of jointly modeling all tweet components in a holistic embedding, which can then be used to tackle various machine learning applications. To address this problem, we propose a deep neural network framework that combines text, image, and graph representations to learn joint embeddings for 5 tweet components: body, hashtags, images, user, and location. In our experiments, we use a large dataset of tweets to learn a joint embedding model and use it in multiple tasks to evaluate its performance vs. state-of-the-art baselines specific to each task. Our results show that our proposed generic method has similar or superior performance to specialized application-specific approaches.

7.1 Introduction

Twitter produces a wealth of information for analysis of trends, opinions, and interactions, with 500 million tweets per day generated by its users [156]. As such, the microblogging service is a popular target for research involving machine learning. Several problem settings in the field of machine learning, or variants thereof, can focus on Twitter data (Figure 28). For example, sentiment analysis, spam detection, and location prediction, all well-established problem settings in their own right, are all applicable to tweets [157-159]. Much of the work in applying machine learning to Twitter data focuses on only one or a few components of a tweet. A sentiment analysis model might only use the text of a tweet, while a hashtag recommendation system might use the text and image. A tweet can contain additional components that may be informative to a machine learning model. Examples of these components include the user, whose interactions with other users can be modeled by a graph, and the location, which can link a tweet to others at the same location. Incorporating several of these tweet components into a machine learning framework can potentially create a model that is better informed than others at a given task. One approach to accomplishing this is by creating a joint embedding framework.



@CapsnYanks713

Nieds retired jersey at prudential

3:55 PM · Jan 3, 2015
from Newark, NJ

@womynrebel

Legal observing as Westchester takes the streets. #BlackLivesMatter #914ShutItDown #MillionsMarch @NYCLU @CandisLynnT

11:06 AM · Dec 13, 2014

@xoxoneil

thank you @god @jesus @BurgerKing #ChickenFriesAreBack ~you are the music in me plays in a distance~

2:50 PM · Aug 11, 2014
from New Jersey, USA

Figure 28. Some machine learning applications for Twitter[25].

Joint embeddings are used in several machine learning tasks that handle different modalities, e.g. text and images, in order to leverage the relationship between them. The intuition behind a joint embedding space is that inputs from different modalities mapped into the space should be close if they are semantically related, e.g. for the problem of image-text retrieval, the image caption or tags closest to an image should be those that best describe the visual content of the image. However, these models are generally limited to 2 or 3 modalities (components), like the aforementioned text and image.

[25] From left: hashtag recommendation, location prediction, sentiment analysis.

Introducing additional modalities has two potential benefits. First, it can better inform existing applications by taking these additional modalities into account. In the image-text retrieval task, also considering the author and the location of a post from an image sharing or microblogging service might achieve better results. Second, introducing additional modalities can open up a joint embedding space to new applications. Recent work in hashtag recommendation uses the text and image of a social media post as input to a neural network model [160-162], but a joint embedding space that includes hashtags as well as text, images, and other modalities, might perform well at this task. Figure 29 shows an illustration of a similar scenario. The first question we ask in this paper within the context of Twitter is, *can additional modalities in a joint embedding space improve its performance in typical applications and/or enable it to perform well in new ones?* The second question is, *can we build a single holistic embedding model for tweets and reuse it in several diverse applications?*

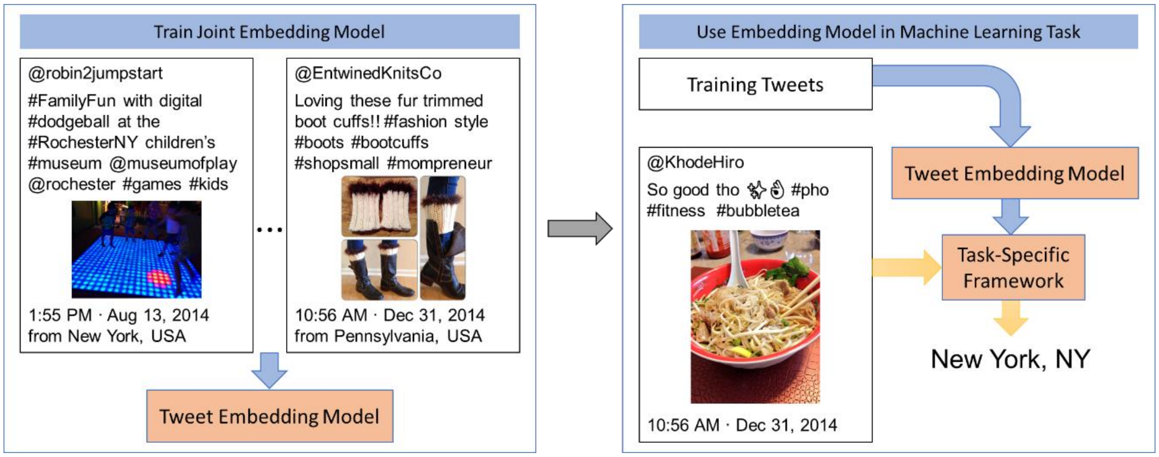

Train Joint Embedding Model
@robin2jumpstart
#FamilyFun with digital #dodgeball at the #RochesterNY children's #museum @museumofplay @rochester #games #kids
1:55 PM · Aug 13, 2014
from New York, USA
…
@EntwinedKnitsCo
Loving these fur trimmed boot cuffs!! #fashion style #boots #bootcuffs #shopsmall #mompreneur
10:56 AM · Dec 31, 2014
from Pennsylvania, USA
Tweet Embedding Model
Use Embedding Model in Machine Learning Task
Training Tweets
@KhodeHiro
So good tho #pho #fitness #bubbletea
10:56 AM · Dec 31, 2014
Tweet Embedding Model
Task-Specific Framework
New York, NY

Figure 29. The problem setting of this paper. Our goal is to learn a joint embedding consisting of multiple tweet components and successfully apply it to several machine learning tasks for Twitter.

To answer these questions, we propose a deep neural network framework based on previous work in image-text retrieval to learn a joint embedding model for Twitter that incorporates multiple tweet components beyond the text and images typical to existing joint embedding models. We evaluate the model's performance on several Twitter machine learning tasks and show that it meets or exceeds the performance of baselines selected from prior work that addresses these tasks.

7.1.1 Overview of the Proposed Approach

Our approach builds on the VSE++ framework proposed by Faghri et al. [163]. VSE++ learns a joint embedding space for cross-modal image-text retrieval, i.e. image captioning and image search. The main contribution of VSE++ is the incorporation of hard negatives into the loss function, which have been shown to be effective in several tasks [164-166]. The structure of the VSE++ network consists of two parallel branches for im-ages and captions. Image features are computed by a convolutional neural network; VGG19 [167] and ResNet152 [168], both pre-trained on the ImageNet dataset [169], are both evaluated in their experiments. Text features from image captions are generated by passing word embeddings to a recurrent neural network; gated recurrent units (GRUs) [170] are used in their experiments.

In this work, we extend the VSE++ framework to incorporate 3 more tweet components in addition to text and images: hashtags, considered separately from tweet text; users, as represented by a graph embedding learned from a graph of Twitter user mentions; and location, to represent the context of a tweet in terms of what other Twitter users are talking about at the same place. Extending VSE++ is not trivial, as adding 3 additional modalities complicates training. The loss function involved in training the model must account for how one modality interacts with four others rather than just one.

Our proposed model is applicable to several tasks, including:

- **Image-text retrieval.** Given a tweet t, predict an image relevant to t, or vice-versa.
- **Hashtag recommendation.** Given a tweet t, predict one or more hashtags relevant to t.
- **Bot detection.** Given a user u and several tweets written by u, predict whether or not u is a bot, i.e. an automated Twitter account.
- **Location prediction.** Given a tweet t, predict where the author of t was when it was posted.

In our experiments, we show the performance of our proposed model compared to baselines from these domains using Twitter data.

7.1.2 Contributions

We make the following contributions:

- We propose an approach that incorporates 5 tweet components to learn a joint embedding model for tweets. This approach extends previous work on image-text retrieval.
- We develop a novel framework with pair-wise ranking loss to learn a robust joint embedding with 5 tweet components.
- We demonstrate that our proposed framework meets or exceeds the performance of baselines in several machine learning tasks.

7.2 Related Work

Joint Embedding: Joint embedding models have been proposed for image-text retrieval [163, 171-173], video-sentence retrieval [174-178], video-paragraph retrieval [179, 180], temporal localization of moments [181-183], and a variety of other tasks [184-187]. The general idea behind a joint embedding model is to place vector representations of different media, such as text and images, into the same embedding space such that the distance between semantically similar vectors (e.g. an image and its captions or tags) is minimized. For the image-text retrieval task, Faghri et al. [163] projected images and text in the visual-semantic embedding space and learned the model utilizing hard negatives. In [171], Mithun et al. used images and noisy text from the Web to improve the joint embedding model. Lee et al. [173] captures fine-grained interplay between objects present in an image and text to better align images and text in the joint embedding space. For the video-sentence retrieval task, Mithun et al. [174] employed multimodal cues such as image, motion, and audio for video encoding. In [175], multi-level encodings for video and text were used and both videos and sentences were encoded in a similar manner. Among the recent works of video-sentence retrieval, Wray et al. [176] enriched embedding learning by disentangling parts-of-speech of captions. For the temporal localization task, moment-sentence pairs [181, 183] or clip-sentence pairs [182] are aligned in the joint embedding space.

Machine Learning on Twitter Data: Several studies have created machine learning methods specifically for use with Twitter data [188-190]. However, these works typically focus on text instead of also including other parts of a tweet. Some methods developed for Twitter and other microblogging services also incorporate images for tasks such as hashtag recommendation [161, 191, 192] and location prediction [193]. However, these approaches ignore other information pre-sent in tweets that a method could otherwise leverage to achieve better results for these and other applications.

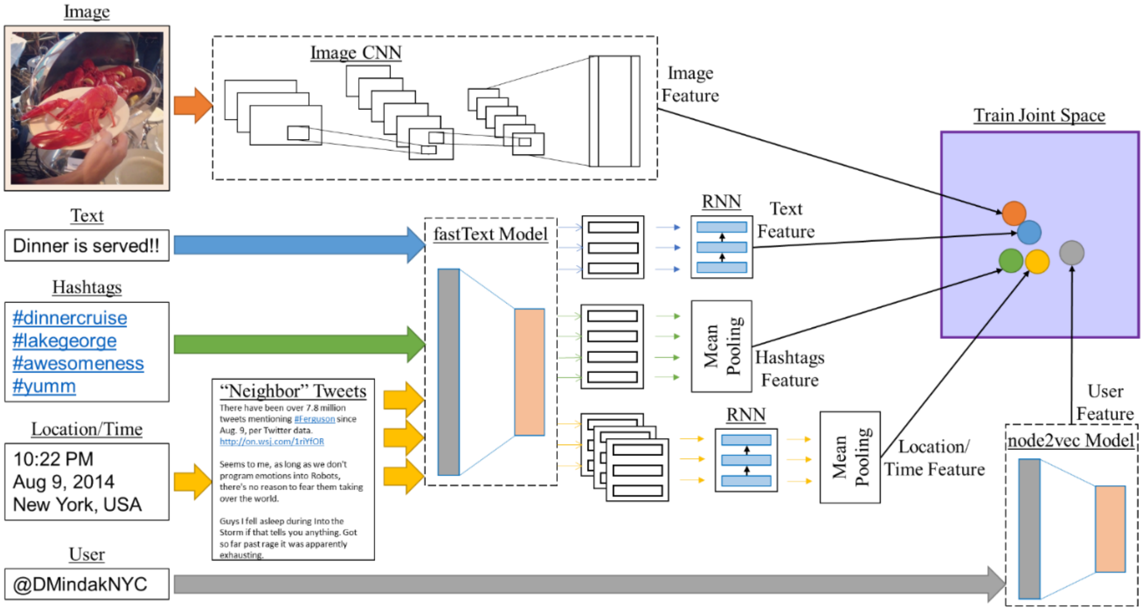
As our proposed model is applicable to several tasks, we discuss further related work when describing the baselines for the tasks evaluated in our experiments in Section 7.4.

7.3 Approach

In this section, we describe our proposed model to represent tweets in an embedding space. We first describe the structure of the neural network framework and how we represent tweet components (Section 7.3.1). Then, we present our approach to training a joint embedding model with this framework using pairwise ranking loss (Section 7.3.2).

7.3.1 Network Structure and Input Features

Network Structure: We learn a joint embedding model using a deep neural network framework. Our framework, shown in Figure 30, has 5 branches for tweet text, an image, hashtags, the location and time of the tweet, and the user who wrote the tweet. Each of these branches uses a different network. As explained in [171], the goal of this design is for the individual branch networks to focus on component-specific features while the fully connected layers convert these to component-robust features.



There have been over 7.8 million tweets mentioning #Ferguson since Aug. 9, per Twitter data. http://on.wsj.com/1riYfOR

Seems to me, as long as we don't program emotions into Robots, there's no reason to fear them taking over the world.

Guys I fell asleep during Into the Storm if that tells you anything. Got so far past rage it was apparently exhausting.

Figure 30. Overview of our proposed framework for learning a multimodal embedding model for tweets. A dataset of tweets is used to learn an aligned representation of tweet components. The trained embedding is used in several tasks.

Text Representation: While we represent hashtags separately from the text, the input to the text branch of our framework also includes the hashtags present in the tweet. The reasoning behind this is that hashtags are often used as words within sentences of tweet text. Pruning the hashtags could thus remove important semantic information.

For encoding tweet text, we use an embedding layer with weights initialized with word embeddings from a fastText [85] model trained on Twitter data. The dimensionality of this layer is 300. The word embeddings are then input to GRUs, which are commonly used to represent sentences [171]. The dimensionality of the GRU output, as well as the joint embedding space *D*, is 1024.

Image Representation: To encode an image contained in a tweet, we use a 152-layer ResNet model [168] trained on the ImageNet dataset [169]. The dimensionality of the image embedding is 2048; this is mapped to the joint space via a fully connected layer.

Similar frameworks have also evaluated a 19-layer VGG model [167] as an alternative, however ResNet has been shown to perform better, at least for the task of image-text retrieval [163, 171], so we limit our experiments to the ResNet model.

Hashtag Representation: To represent hashtags, we separate hashtags from the text of the tweet and average over the fastText word embeddings of those hashtags then map this to the joint space with a fully connected layer.

Location/Time Representation: To emphasize the context of a tweet, we consider time and location in terms of tweets from the same time or place, respectively. This is a two-step process: first, we collect the text of "neighbor" tweets from the same time or place as an input tweet. For simplicity, we do this by grouping tweets from our collected Twitter data into one-hour blocks (for time) and grouping tweets with the same Twitter-assigned place ID as the input tweet (for location). The texts of these tweets are then encoded through network branches identical to that of the input tweet's text, but this is followed by averaging the encodings of the texts.

In our early experiments, we found that retrieval of time embeddings performed poorly, so we exclude time representation from our experimental evaluation.

User Representation: We represent the author of a tweet by using a trained graph embedding model. Specifically, we use the fastnode2vec [194] implementation of node2vec [195]. The weighted graph used to train this model was constructed with users as vertices and mentions as edges, i.e. an edge (u, v, w) represents user u mentioning user v in w tweets. The dimensionality of the graph embedding is 300; this is mapped to the joint space via a fully connected layer.

7.3.2 Training Joint Embedding

For a tweet t, other tweets that are not similar to t should have embedding vectors for their components that are not similar to the embedding vectors of the components of t. With that intuition in mind, our goal is to learn a joint embedding characterized by the weights of the fully connected layers, the text and location/time word embedding layers, and the GRUs.

We base our approach on previous work that uses hinge-based bi-directional ranking loss for visual-semantic embeddings [163, 174]. These approaches maximize the similarity between corresponding image and text embeddings and minimize similarity to non-matching embeddings. They also focus on hard negatives, i.e. given a pair (i, t) of image and text embedding vectors, the corresponding hard negatives are the image vector $\hat{\imath} \neq i$ and the text vector $\hat{t} \neq t$ closest to t and i, respectively.

Our approach must also account for hashtags, author, and location. To accomplish this, we first calculate the loss using each pair (c, a), where c is the embedding for one component from a tweet (e.g. text or image) and a is the averaged tweet component embeddings from the same tweet. This can be written as follows:

$$\begin{aligned}\mathcal{L}_{ca} = \sum_{(c,a)} \{&max[0, \Delta - f(c,a) + f(c,\hat{a})] \\ &+ max[0, \Delta - f(a,c) + f(a,\hat{c})]\end{aligned} \qquad (11)$$

where Δ is the margin value for the ranking loss, $f(c,a) = f(a,c)$ is the similarity scoring function between a tweet component embedding c and averaged tweet component embeddings a, and $\hat{a}$ = $\arg\max_{a^-} f(c, a^-)$ and $\hat{c}$ = $\arg\max_{c^-} f(a, c^-)$ are the hardest negative samples. In our experiments we use cosine similarity for $f(c, a)$, but our approach does not depend specifically on this. With Equation 11 and the set of tweet component embeddings (t, i, h, l, u), representing text, image, hashtags, location, and user of a tweet, respectively, our complete loss function is

$$\mathcal{L} = \mathcal{L}_{ta} + \mathcal{L}_{ia} + \mathcal{L}_{ha} + \mathcal{L}_{la} + \mathcal{L}_{ua} \quad (12)$$

i.e. the total loss is the sum of the loss calculated using one tweet component and the averaged tweet components.

7.4 Experiments

Our experiments demonstrate the proposed model's effectiveness on several machine learning applications involving Twitter data by comparing the results of experiments versus baselines that address those applications. In each of these experiments, we generate tweet component embeddings from our trained model and use them in an application-specific framework as shown in Figure 31. Note that the tweet component embedding model is only trained once, then used in all of the applications evaluated in our experiments. Table 62 summarizes the applications, baselines, and performance metrics in our experiments.

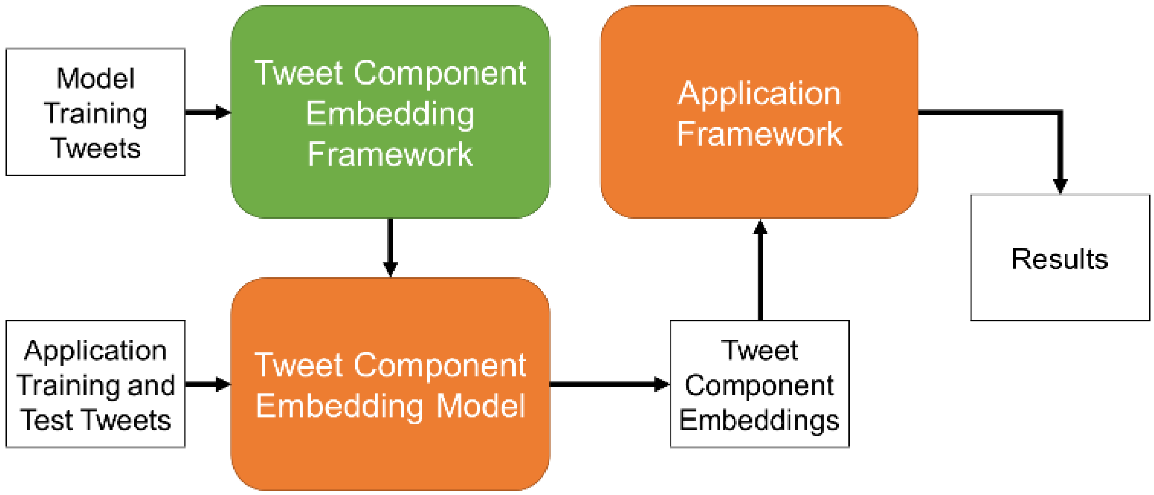

Model Training Tweets
Tweet Component Embedding Framework
Application Framework
Results
Application Training and Test Tweets
Tweet Component Embedding Model
Tweet Component Embeddings

Figure 31. Overview of application-specific experiments. Our trained tweet component embedding model generates embeddings to be used in an application-specific framework.

Table 62. Applications and baselines evaluated.

Section	**Application**	**Baseline(s)**	**Performance Metrics**
7.4.3	Image-text retrieval	VSE++ [163]	Recall @ *k*, median rank
7.4.4	Hashtag recommendation	Co-Attention [161]	Precision @ *k*, recall @ *k*, F_1 score @ *k*
7.4.5	Bot detection	Botometer v4 [196]	F_1 score, AUC
7.4.6	Location prediction	Deepgeo [197]	Accuracy

7.4.1 Dataset

We trained our model on a dataset of tweets to apply it to each machine learning task. Each tweet in the dataset contains all of the components necessary to generate embedding vectors with our proposed model (i.e. text, image, hashtags, geolocation data, and author ID). 100,000 tweets in the dataset from March 1-8, 2020 are used for training, while 5000 tweets from March 9, 2020 are used for validation. An additional 5000 tweets from March 10, 2020 are used for testing. The tweets in the dataset are limited to those with a location that falls within a bounding box that encompasses most of North America.

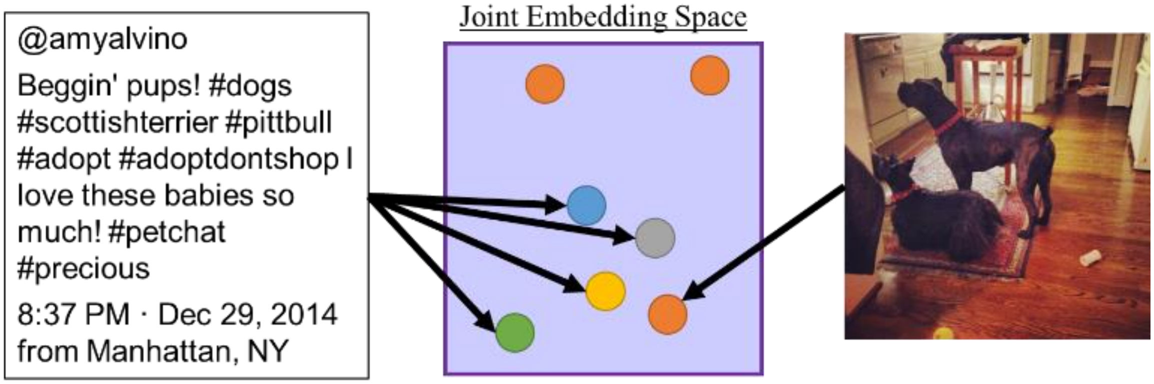
7.4.2 Training Details

The embedding networks in our model are trained with an Adam optimizer [198] over a total of 30 epochs. We set the initial learning rate of 0.0002 and decrease the learning rate by a factor of 10 after 15 epochs. The gradient L2 norm threshold for clipping gradients is 2. The margin Δ is 0.2. We use a mini-batch size of 128. The model is evaluated on the validation set every 500 training iterations. The trained model used for evaluation on the test data is selected based on the sum of recalls (recall @ 1, 5, and 10) on the validation set to mitigate overfitting. The fastText and node2vec models used in our proposed model were trained on tweets from March 1-7, 2020.

7.4.3 Image-Text Retrieval

For image-text retrieval, we compare our proposed model to VSE++ [163], on which our method is based. We also include results from using hashtags as the text for VSE++ because many hashtags have significant descriptive information of their associated images [199], which the text of a tweet might not. In our experiments, which are based on the experiments in [163], each model is given a test tweet minus the image (or text) and attempts to retrieve a relevant image (or text) from the test data.

While retrieving an image with VSE++ is simply a matter of finding the most similar image to the input text (or vice versa), this becomes somewhat more complex with the additional embeddings in our proposed model's joint embedding space. To determine which image (or text) embedding to retrieve, we retrieve the embedding corresponding to the highest similarity score from any of the input tweet's component embeddings. This is shown with our proposed model in Figure 32, where the image corresponding to the image embedding closest to the input tweet's component embeddings is shown.


@amyalvino

Beggin' pups! #dogs #scottishterrier #pittbull #adopt #adoptdontshop I love these babies so much! #petchat #precious

8:37 PM · Dec 29, 2014
from Manhattan, NY

Figure 32. Image retrieval task. A tweet's components are calculated in the joint embedding space and the most similar image is determined.

Our results are shown in Table 63 (image retrieval) and Table 64 (text retrieval), which show that our method outperforms VSE++ in image-text retrieval of Twitter data. However, we note that these findings do not extend to the more general problem of image-text retrieval, as a tweet's text and other non-image components may not be as semantically similar to its image as a caption written specifically for that image. This is supported by the image retrieval results, where we observe that both methods perform very poorly.

Table 63. Image retrieval results.

Method	Recall@1	Recall@5	Recall@10	Median Rank
VSE++	0.0002	0.001	0.002	2500
VSE++ with hashtags as text	0.0002	0.0016	0.0034	2329
Proposed method	0.0004	0.0024	0.0036	2213

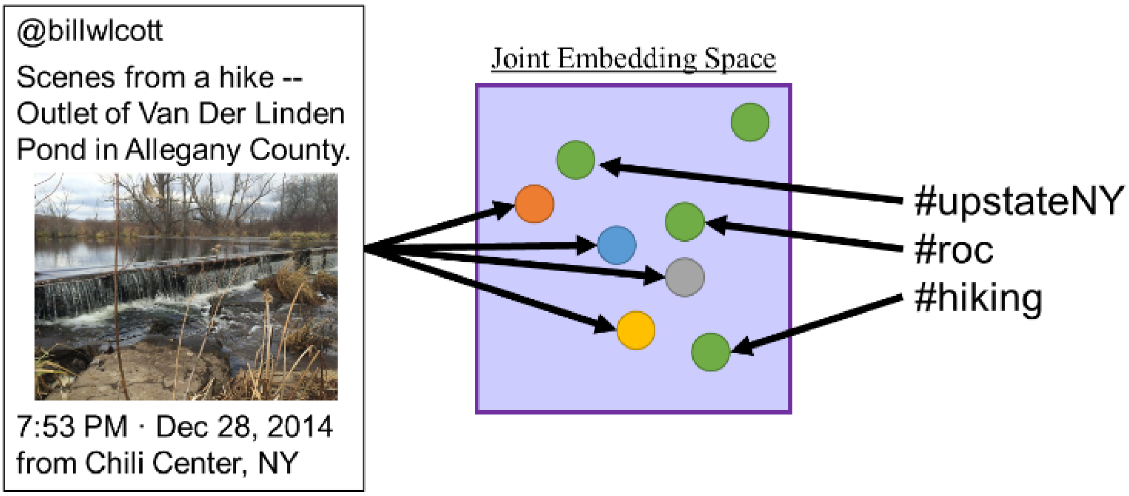
Table 64. Text retrieval results.

Method	Recall@1	Recall@5	Recall@10	Median Rank
VSE++	0.0002	0.001	0.0018	2492
VSE++ with hashtags as text	0.0002	0.0008	0.0016	2481
Proposed method	0.1740	0.2504	0.2820	308

7.4.4 Hashtag Recommendation

Our experiments on hashtag recommendation are performed in the context of recommending hashtags for images and their associated text. Our primary baseline for this is the Co-Attention model proposed in [161]. The Co-Attention model uses both the image and the text of a photo sharing service such as Twitter or Instagram in a deep neural framework that includes a co-attention mechanism [200] to model the interaction between the input image and text. We trained the baseline on the same dataset used to train our model described in Section 7.4.1.

To recommend hashtags for a test tweet t with our model, we use a nearest neighbor-style approach by calculating the embeddings of the components of t and scoring each training hashtag embedding h according to the maximum cosine similarity between h and the component embeddings of t. The recommendation of k hashtags for t is thus the top k hashtags according to these scores. This approach is illustrated in Figure 33 for k = 3, in which the 3 hashtags corresponding to the 3 hashtag embeddings closest to any component embedding from the input tweet are returned.



@billwlcott

Scenes from a hike -- Outlet of Van Der Linden Pond in Allegany County.

7:53 PM · Dec 28, 2014
from Chili Center, NY

Figure 33. Hashtag recommendation task. A tweet is given as input to the hashtag recommendation framework, which then out-puts a number of recommended hashtags.

Figure 34 shows the average precision, recall, and F_1 scores at 1, 5, and 10 for both our method and the Co-Attention baseline. Our method performs better in all three measures for all values of k evaluated.

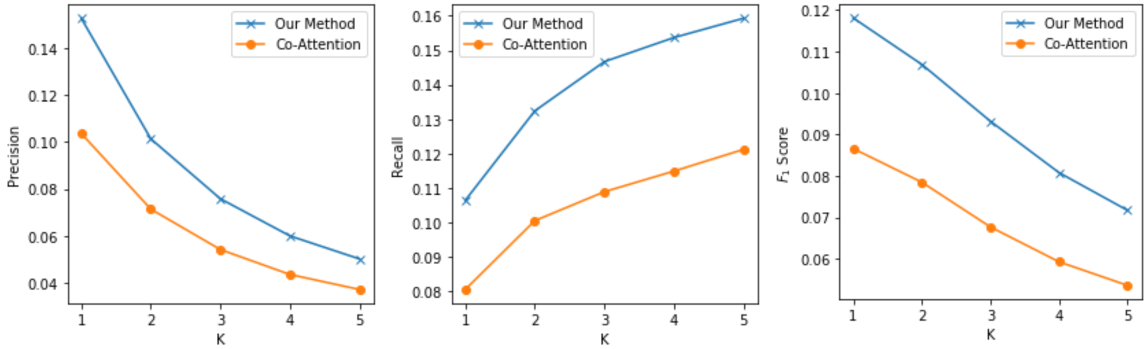


Figure 34. Hashtag recommendation results: precision, recall, and F_1 score with different numbers of recommended hashtags.

7.4.5 Bot Detection

The state-of-the art for bot detection on Twitter is Botometer [80]. Its most recent version, v4 [196], is an ensemble classifier that combines Botometer v3 [201] with several random forest classifiers [82] that are each trained on a specific class of Twitter bot. Botometer v4 serves as our baseline in these experiments. As we use the same datasets in our experiments as those used in [196], we compare the results of our method to those presented there. Specifically, we compare to their cross-domain experiments, which combine several annotated bot detection datasets [196, 201-209] for a training dataset of 43,576 bots and 32,849 humans and a test dataset of 9,432 bots and 8,862 humans. These datasets consist of a Twitter user ID combined with a binary class label indicating whether the user is a bot or a human.

We evaluated an approach that uses tweet component embeddings in a convolutional neural network classifier. For each user in the bot detection datasets, we retrieve up to

200 tweets. Each user is represented by up to n of the most recent of these tweets, where n is a hyperparameter of our bot detection framework. In our experiments, we use n = 10. The component embeddings of each of these tweets is then computed by our model and averaged; each instance is thus an $n_u \times m$ matrix composed of $n_u \leq n$ averaged component embeddings of length m of tweets from user u. The network's structure is based on the text CNN proposed by Kim [84], except the input is matrices of averaged embedding vectors as described above rather than matrices of word embedding vectors. In addition to representing each user by up to n consecutive tweets, we also attempt to balance the training data by adding duplicates of users of the less frequent class (humans) with a different set of up to n tweets drawn from that user's retrieved set of tweets. Training the CNN is performed via 5-fold cross-validation of the training data. This approach is demonstrated in Figure 35, which shows the conversion of a tweet's components to embedding vectors, which are then averaged. The user's averaged tweet embeddings are used as input to a CNN classifier, which predicts the user's class label.

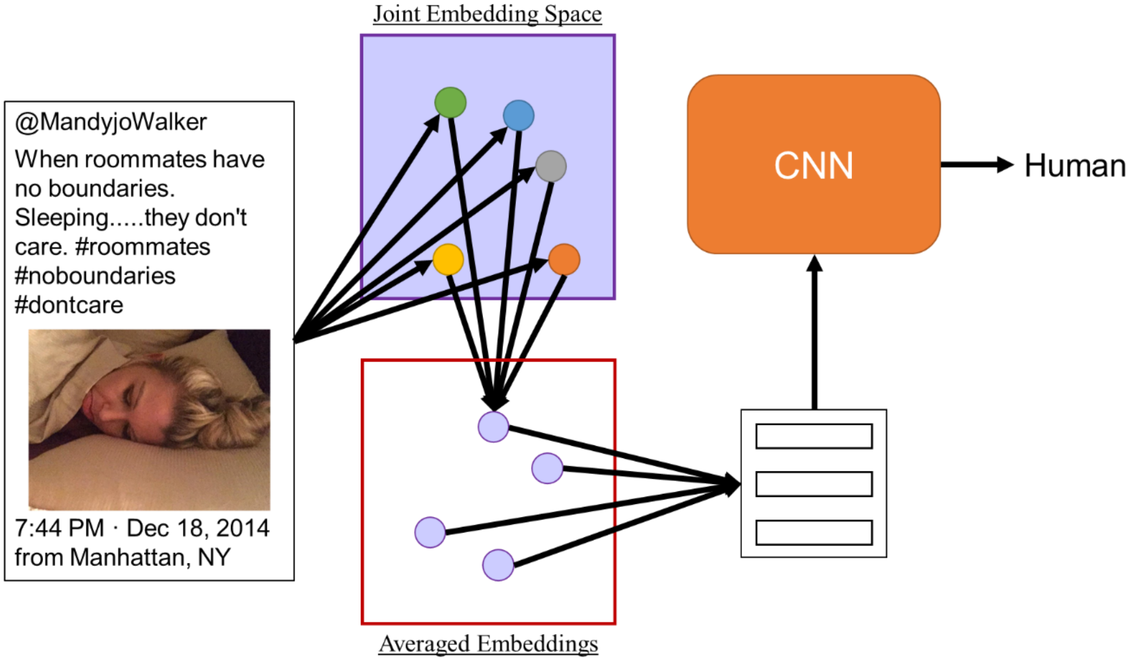


Figure 35. Bot detection task. Component embeddings from a tweet are averaged; a matrix of these vectors from the same user is input to a CNN classifier to predict the user's label.

We further refined our approach by using a validation dataset taken from the training data to tune some of its hyperparameters using a small set of values for each one. Based on these results, we set the CNN's filter window sizes to 2, 3, and 4, and number of feature maps to 300.

The results of our bot detection experiments are shown in Figure 36, which shows the F_1 and AUC scores for the baseline and our proposed approach for the combined test dataset as well as individual bot detection datasets. The baseline's performance is better for all datasets, however our proposed method comes close in some cases. One limitation of our method is the data used to train the tweet component embedding model. Our methods may perform better in the bot detection task if the model were trained with tweets from users in the bot detection training data. In the dataset described in Section 7.4.1, bots may be underrepresented compared to the bot detection data because bots are estimated to make up only 9-15% of active Twitter accounts [203]. Another possible reason for the poor performance of our method compared to the baseline is that the baseline takes advantage of some manual work in separating bots in the training datasets according to distinct bot classes; their model takes advantage of this additional information while our method only considers the binary classes of "bot" and "human." Addressing these issues may improve the performance of our bot detection method.

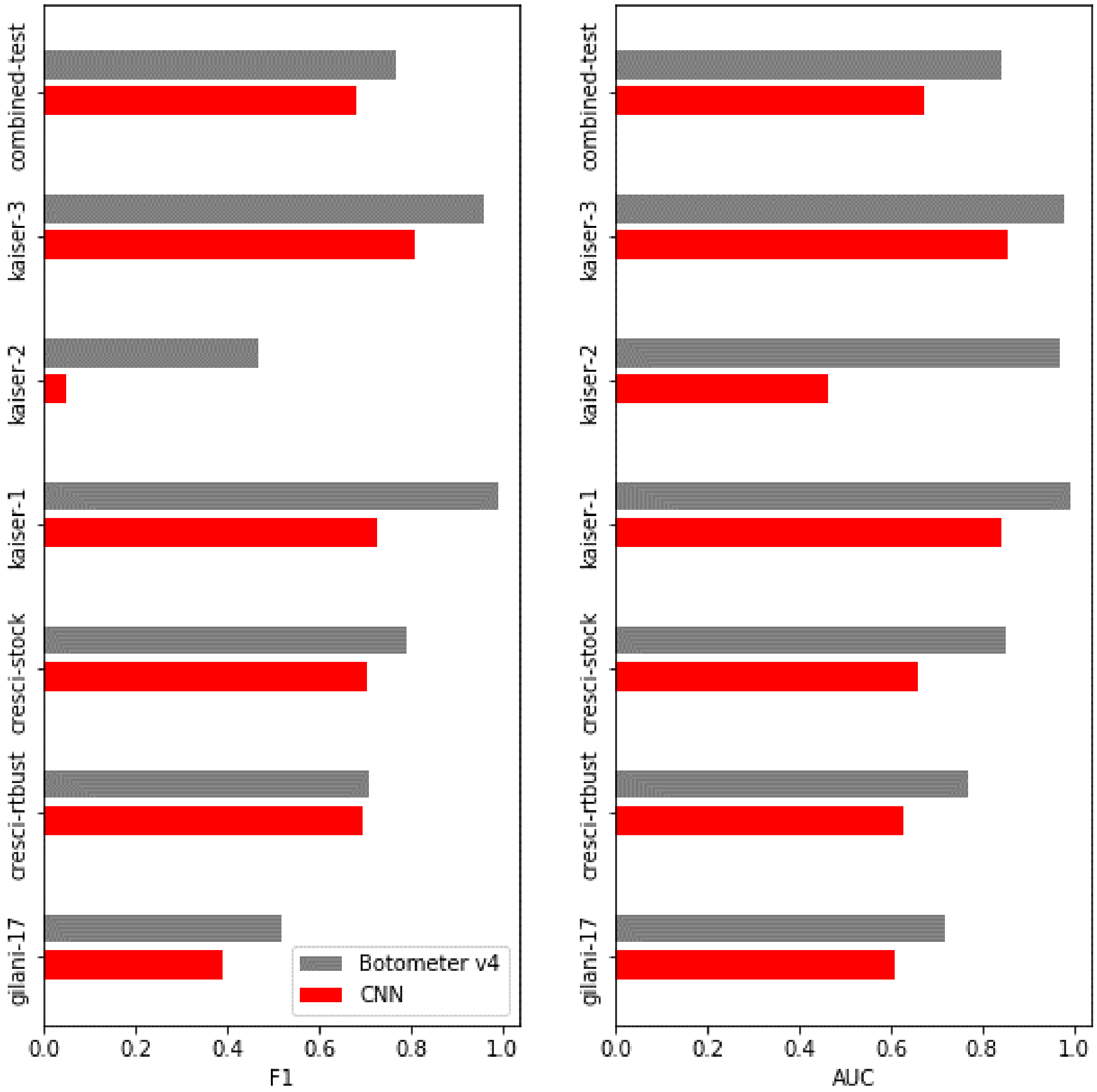
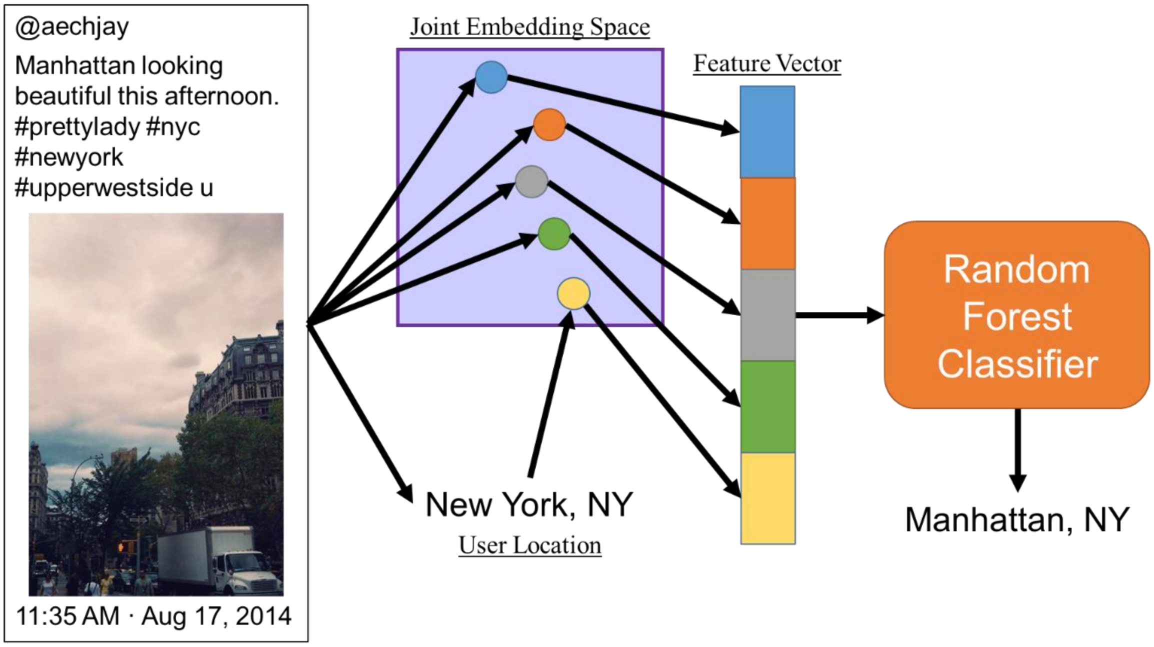


Figure 36. Bot detection results: F_1 score and AUC on various bot detection test datasets.

7.4.6 Location Prediction

We evaluate our proposed model on the task of location prediction in terms of cities, i.e. given a tweet without location data, predict the city where that tweet was posted. As a baseline, we use Deepgeo [197], which takes a tweet's text and creation time as well as the author's UTC offset, time zone, location (i.e. the free text location listed in a user's profile), and account creation time as input to a deep learning framework. Notably, Deepgeo is designed for the classification setting of predicting a city class, rather than predicting a location's latitude and longitude. To train and test this baseline, we used subsets of the datasets described in Section 7.4.1 that include only tweets with a Twitter place ID corresponding to a city (i.e. we omitted tweets with a place ID corresponding to other place types such as administrative regions and points of interest) that is present in all datasets. This left a total of 41,849 tweets in the training data, 2841 tweets in the validation data, and 2594 tweets in the test data, with a set of 522 classes (cities) between them.

Our approach for using our proposed tweet embedding model uses concatenated embeddings, i.e. each tweet is represented by a single feature vector that contains each of the tweet's non-location component embeddings. Noting that Lau et al. [197] found that a user's location contributed substantially to their tweet location prediction model's accuracy, we also represent user location in the tweet's feature vector. We do so by using the user location text of each tweet's author as input to the text branch of our trained tweet component embedding model and concatenating the resulting embedding with its corresponding tweet's component embeddings. We then use these feature vectors and their corresponding class labels as input to a random forest classifier. We use the default hyperparameter values defined in the implementation in [86] with the exception of using “balanced subsample” class weights, e.g. class weights are inversely proportional to class frequencies for every decision tree’s bootstrap sample, to account for any differences in class frequencies within the training data. Our tweet location prediction approach is summarized in Figure 37, where the embeddings of a tweet’s non-location components and its author’s location text are computed and concatenated to a single feature vector, then passed to a random forest classifier to predict the tweet’s location.



@aechjay

Manhattan looking beautiful this afternoon. #prettylady #nyc #newyork #upperwestside u

11:35 AM · Aug 17, 2014

Figure 37. Location prediction task. The input tweet’s component and user location embeddings are concatenated and input to a classifier that predicts the tweet’s location.

The results of our location prediction experiments are shown in Table 65, which show that our method has higher accuracy than the Deepgeo baseline.

Table 65. Location prediction results.

Method	**Accuracy**
Deepgeo	48.88%
Proposed method	52.43%

7.5 Conclusions

In this paper, we proposed a joint embedding framework for representing multimodal tweet data to generate embeddings for machine learning tasks on Twitter. The framework aligns tweet component embeddings in the joint space using a loss function that incorporates hard negatives. We tested a trained tweet component embedding model on four Twitter machine learning applications and found that it can perform well on most applications evaluated.

Chapter 8

Conclusion

This dissertation presented analyses of Web content and also introduced new machine learning methods to analyze user-generated content on the Web.

In Chapter 2, we compared health insurance plan quality measures to attributes of health care providers within those plans' networks. We found that insurance plan consumer satisfaction is positively correlated with both patient ratings and relative costs of health care providers, but many other correlations were negligible. These findings may provide new insights to help patients, insurers, and health care providers alike.

In Chapter 3, we analyzed real estate prices near universities and hospitals. This analysis was conducted both in terms of median ZIP code prices and the prices of individual homes. Among our findings were that ZIP codes with a university tend to have higher prices, ZIP codes with a smaller hospital tend to have lower prices, and smaller homes tend to be more sensitive to distance from a university. Our results showed some of the complexities of real estate economics but may contribute to a machine learning model for real estate prices in the future.

In Chapter 4, we surveyed the frequencies of health-related post content categories in terms of user demographics from both general social networks and health-related Web forums. We found that male users asked for medical advice on WebMD more frequently than female users, sharing experiences is popular with every demographic on

DailyStrength, all demographics share experiences more often than they share news on Twitter, and educational material is shared least frequently by users age 35-44 and most frequently by Asian users. Our findings may help researchers and health advocates reach the demographics they seek for clinical trials and information campaigns.

In Chapter 5, we presented the doctor review classification problem, in which a review sentence may represent one of two opposing opinions on some aspect of a doctor visit, or may be unrelated to that aspect. With a dataset of doctor review sentences labeled with several opinion classes, we evaluated several classification methods, including a new NLP-based method, and demonstrated the feasibility of addressing this problem.

In Chapter 6, we introduced the problem of retrieving an effective dataset for training a binary social media post classifier when constrained by a search interface. We discussed this problem in terms of achieving both balance, i.e. a dataset with a number of positives and negatives as close to equal as possible, and diversity, i.e. a wide range of samples for both the positive and negative classes. Our proposed method retrieved a dataset that was able to train a classifier with higher accuracy than several baselines.

In Chapter 7, we built upon prior work in image-text retrieval to create a joint embedding model for Twitter that takes advantage of several tweet components. Within the context of Twitter, our framework is applicable to existing problems addressed by joint embeddings but may be applied to other tasks as well. Our results show that our proposed model meets or exceeds the performance of baseline methods for some of these tasks.

Bibliography

[1] Weiss A. Health Affairs Blog. 2012 Jul 26. Health Insurance Exchanges: Improving Health Care Access and Quality. URL: http://healthaffairs.org/blog/2012/07/26/health-insurance-exchanges-improving-health-care-access-and-quality/.

[2] Castle Connolly Top Doctors. URL: https://www.castleconnolly.com/.

[3] NCQA. URL: https://www.ncqa.org.

[4] EINSURANCE. Privacy, Trust, and User Data. URL: http://www.einsurance.com/info/privacy/.

[5] Insure.com. Methodology & Disclaimer. URL: http://www.insure.com/best-insurance-companies/methodology.

[6] Feldman PH. Center for Health Care Strategies. 2014 Aug. Key Attributes of High-Performing Integrated Health Plans for Medicare-Medicaid Enrollees. URL: http://www.chcs.org/media/PRIDE-Key-Attributes-of-High-Performing-Health-Plans_090514.pdf.

[7] URAC. Health Plan Quality Measures. URL: https://www.urac.org/accreditation-and-measurement/accreditation-programs/all-programs/health-plan-quality-measures/.

[8] Vitals. URL: http://www.vitals.com/.

[9] Healthgrades. URL: https://www.healthgrades.com/.

[10] U.S. News & World Report. Best Hospitals: National Rankings. URL: http://health.usnews.com/best-hospitals/rankings.

[11] U.S. News & World Report. Best Medical Schools: Primary Care. URL: http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-medical-schools/primary-care-rankings.

[12] HealthData.gov. National Provider Identifier Standard - Data Dissemination. URL: http://www.healthdata.gov/dataset/national-provider-identifier-standard-data-dissemination.

[13] Medicare.gov. Physician Compare Datasets. URL: https://data.medicare.gov/data/physician-compare.

[14] Centers for Medicare & Medicaid Services. Medicare Provider Utilization and Payment Data: Physician and Other Supplier. URL: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html.

[15] Centers for Medicare & Medicaid Services. What physician shared patient data sets are available? URL: https://questions.cms.gov/faq.php?faqId=7977.

[16] Centers for Medicare & Medicaid Services. Part D Claims Data. URL: http://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/PartDData.html.

[17] Hedley J. jsoup: Java HTML Parser. URL: https://jsoup.org/.

[18] Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 1966;10(8):707-710.

[19] Washington Publishing Company. WPC References. URL: http://www.wpc-edi.com/reference/.

[20] Dolan PL. American Medical News. Physician rating website reveals formula for good reviews. URL: http://www.amednews.com/article/20120227/business/302279969/2/.

[21] Gao GG, McCullough JS, Agarwal R, Jha AK. A changing landscape of physician quality reporting: Analysis of patients' online ratings of their physicians over a 5-year period. J Med Internet Res 2012 Feb 24;14(1):e38.

[22] Bostic RW, Longhofer SD, Redfearn CL. Land leverage: Decomposing home price dynamics. Real Estate Econ. 2007;35(2):183-208.

[23] Diewert WE, Haan JD, Hendricks R. Hedonic regressions and the decomposition of a house price index into land and structure components. Econometric Rev. 2015;34(1-2):106-126.

[24] Benson ED, Hansen JL, Schwartz AL, Smersh GT. Pricing residential amenities: The value of a view. J Real Estate Finance Econ. 1998;16(1):55-73.

[25] van Praag B, Baarsma BE. The shadow price of aircraft noise nuisance (No. 01-010/3). Tinbergen Institute Discussion Paper. 2001.

[26] Ozdenerol E, Huang Y, Javadnejad F, Antipova A. The impact of traffic noise on housing values. J Real Estate Pract Educ. 2015;18(1):35-54.

[27] Barr JR, Ellis EA, Kassab A, Redfearn CL, Srinivasan NN, Voris KB. Home price index: A machine learning methodology. Int J Semantic Comput. 2017; 11(01):111-133.

[28] Cesa-Bianchi A, Cespedes LF, Rebucci A. Global liquidity, house prices, and the macroeconomy: Evidence from advanced and emerging economies. J Money Credit Banking. 2015;47(S1):301-335.

[29] Favara G, Imbs J. Credit supply and the price of housing. Am Econ Rev. 2015;105(3):958-992.

[30] Muehlenbachs L, Spiller E, Timmins C. The housing market impacts of shale gas development. Am Econ Rev. 2015;105(12):3633-3659.

[31] Waddell P, Berry BJ, Hoch I. Residential property values in a multinodal urban area: New evidence on the implicit price of location. J Real Estate Finan Econ. 1993;7(2):117-41.

[32] Nau C, Bishai D. Green pastures: Do US real estate prices respond to population health? Health Place. 2018;49(1):59-67.

[33] Otto P, Schmid W. Spatiotemporal analysis of German real-estate prices. Ann Reg Sci. 2018;60(1):41-72.

[34] Rascoff S, Humphries S. Zillow talk: The new rules of real estate. New York: Grand Central Publishing; 2015.

[35] Turner J. The impact of walkability on home values: Findings from neighborhoods in three Bay Area cities. Digital Commons @ Cal Poly. 2017. URL: http://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1178&context=crpsp.

[36] Bolitzer B, Netusil NR. The impact of open spaces on property values in Portland, Oregon. J Environ Manag. 2000;59(3):185-193.

[37] Anderson ST, West SE. Open space, residential property values, and spatial context. Reg Sci Urban Econ. 2006;36(6):773-789.

[38] Debrezion G, Pels E, Rietveld P. The impact of rail transport on real estate prices: an empirical analysis of the Dutch housing market. Urban Stud. 2011;48(5): 997-1015.

[39] Moore CL, Sufrin SC. The impact of a nonprofit institution on regional income. Growth and Change. 1974;5(1):36-40.

[40] Beeson P, Montgomery EB. The effects of colleges and universities on local labor markets. Massachusetts: National Bureau of Economic Research; 1990.

[41] Hedrick DW, Henson ST, Mack RS. The effects of universities on local retail, service, and F.I.R.E. employment: Some cross-sectional evidence. Growth Change. 1990;21(3):9-20.

[42] Moore GA. Local income generation and regional income redistribution in a system of public higher education. J High Educ. 1979;50(3):334-348.

[43] Zillow data. URL: http://www.zillow.com/research/data/.

[44] United States Census Bureau data. URL: http://www.census.gov/data.html.

[45] Wikipedia. URL: http://en.wikipedia.org.

[46] National university rankings. URL: http://www.usnews.com/best-colleges/rankings/national-universities.

[47] Hospital general information. URL: http://data.medicare.gov/Hospital-Compare/Hospital-General-Information/xubh-q36u.

[48] Physician compare national downloadable file. URL: http://data.medicare.gov/Physician-Compare/Physician-Compare-National-Downloadable-File/mj5m-pzi6.

[49] HUD aggregated USPS administrative data on address vacancies. URL: http://www.huduser.gov/portal/datasets/usps.html.

[50] Geographic terms and concepts - census tract. URL: http://www.census.gov/geo/reference/gtc/gtc_ct.html.

[51] Pearson K. Notes on regression and inheritance in the case of two parents. Proc R Soc London. 1895;58(1):240-242.

[52] Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4): e85.

[53] Kane GC, Fichman RG, Gallaugher J, Glaser J. Community relations 2.0. Harv Bus Rev 2009 Nov;87(11):45-50, 132.

[54] Hackworth BA, Kunz MB. Health care and social media: Building relationships via social networks. Acad Health Care Manag J 2011;7(2):1-14.

[55] Wiley MT, Jin C, Hristidis V, Esterling KM. Pharmaceutical drugs chatter on Online Social Networks. J Biomed Inform 2014 Jun;49:245-254.

[56] Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, et al. Psychological language on Twitter predicts county-level heart disease mortality. Psychol Sci 2015 Feb;26(2):159-169.

[57] Yu B, Gerido L, He Z. Exploring text classification of social support in online health communities for people who are D/deaf and hard of hearing. Proc Assoc Info Sci Tech 2017;54(1):840-841.

[58] Reavley NJ, Pilkington PD. Use of Twitter to monitor attitudes toward depression and schizophrenia: An exploratory study. PeerJ 2014;2:e647.

[59] Lee JL, DeCamp M, Dredze M, Chisolm MS, Berger ZD. What are health-related users tweeting? A qualitative content analysis of health-related users and their messages on twitter. J Med Internet Res 2014 Oct 15;16(10):e237.

[60] Lopes CT, da Silva BG. A classification scheme for analyses of messages exchanged in online health forums. Inf Res 2019;24(1).

[61] Krueger PM, Tran MK, Hummer RA, Chang VW. Mortality attributable to low levels of education in the United States. PLoS One 2015;10(7):e0131809.

[62] Anderson-Bill ES, Winett RA, Wojcik JR. Social cognitive determinants of nutrition and physical activity among web-health users enrolling in an online intervention: The influence of social support, self-efficacy, outcome expectations, and self-regulation. J Med Internet Res 2011 Mar 17;13(1):e28.

[63] Sadah SA, Shahbazi M, Wiley MT, Hristidis V. A study of the demographics of Web-based health-related social media users. J Med Internet Res 2015 Aug 6;17(8):e194.

[64] Sadah SA, Shahbazi M, Wiley MT, Hristidis V. Demographic-based content analysis of web-based health-related social media. J Med Internet Res 2016 Jun 13;18(6):e148.

[65] Sadilek A, Kautz H, Silenzio V. Modeling spread of disease from social interactions. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media; 2012 Jun 4.

[66] Huh J, Yetisgen-Yildiz M, Pratt W. Text classification for assisting moderators in online health communities. J Biomed Inform 2013 Dec;46(6):998-1005.

[67] Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015 May;22(3):671-681.

[68] Mislove A, Lehmann S, Ahn YY, Onnela J, Rosenquist JN. Understanding the demographics of Twitter users. In: Proceedings of the 5th international AAAI conference on weblogs and social media; 2011 Jul 17.

[69] Kanthawala S, Vermeesch A, Given B, Huh J. Answers to health questions: Internet search results versus online health community responses. J Med Internet Res 2016 Apr 28;18(4):e95.

[70] Bissoyi S, Mishra BK, Patra MR. Recommender systems in a patient centric social network - a survey. In: Proceedings of the 2016 International Conference on Signal Processing, Communication, Power and Embedded System; 2016 Oct 3; p. 386-389.

[71] RxList - The Internet Drug Index for Prescription Drug Information, Interactions, and Side Effects. URL: http://www.rxlist.com/script/main/hp.asp.

[72] Twitter Developer. URL: https://developer.twitter.com/.

[73] Google+ API | Google+ Platform for Web. | Google Developers. URL: https://www.webcitation.org/78S7b05G0.

[74] Twitter. URL: https://twitter.com/.

[75] Google Plus. URL: https://plus.google.com/.

[76] DailyStrength: Support Groups. URL: http://www.dailystrength.org/support-groups.

[77] WebMD - Better Information. Better Health. URL: http://www.webmd.com/.

[78] Robillard JM, Johnson TW, Hennessey C, Beattie BL, Illes J. Aging 2.0: Health information about dementia on Twitter. PLoS One 2013;8(7):e69861.

[79] Alvaro N, Conway M, Doan S, Lofi C, Overington J, Collier N. Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. J Biomed Inform 2015 Dec;58:280-287.

[80] Davis CA, Varol O, Ferrara E, Flammini A, Menczer F. BotOrNot: A system to evaluate social bots. In: Proceedings of the 25th International Conference Companion on World Wide Web; 2016 Apr 11; p. 273-274.

[81] Hayati P, Chai K, Potdar V, Talevski A. Behaviour-based Web spambot detection by utilising action time and action frequency. In: Proceedings of the International Conference on Computational Science and Its Applications; 2010 Mar 23; p. 351-360.

[82] Breiman L. Random forests. Mach Learn 2011;45(1):5-32.

[83] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273-297.

[84] Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing; 2014 Oct 25; p. 1746-1751.

[85] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017;5:135-146.

[86] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825-2830.

[87] Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw 2018 Oct;106:249-259.

[88] Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: Proceedings of the 20th International Conference on Pattern Recognition; 2010 Aug 23; p. 3121-3124.

[89] Sillence E, Briggs P, Harris P, Fishwick L. Going online for health advice: Changes in usage and trust practices over the last five years. Interact Comput 2007;19(3):397-406.

[90] Heaivilin N, Gerbert B, Page JE, Gibbs JL. Public health surveillance of dental pain via Twitter. J Dent Res 2011 Sep;90(9):1047-1051.

[91] Broniatowski DA, Jamison AM, Qi S, AlKulaib L, Chen T, Benton A, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. Am J Public Health 2018 Oct;108(10):1378-1384.

[92] Yuan X, Schuchard RJ, Crooks AT. Examining emergent communities and social bots within the polarized online vaccination debate in Twitter. Soc Media Soc 2019;5(3):205630511986546.

[93] Mathai E, Allegranzi B, Kilpatrick C, Bagheri Nejad S, Graafmans W, Pittet D. Promoting hand hygiene in healthcare through national/subnational campaigns. J Hosp Infect 2011 Apr;77(4):294-298.

[94] Ding X, Liu B, Yu PS. A holistic lexicon-based approach to opinion mining. In: Proceedings of the 2008 International Conference on Web Search and Data Mining; 2008 Feb 11; p. 231-240.

[95] Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2004 Aug 22; p. 168-177.

[96] Popescu AM, Etzioni O. Extracting product features and opinions from reviews. In: Natural Language Processing and Text Mining. London: Springer; 2007:9-28.

[97] Segal J, Sacopulos M, Sheets V, Thurston I, Brooks K, Puccia R. Online doctor reviews: Do they track surgeon volume, a proxy for quality of care? J Med Internet Res 2012 Apr 10;14(2):e50.

[98] López A, Detz A, Ratanawongsa N, Sarkar U. What patients say about their doctors online: A qualitative content analysis. J Gen Intern Med 2012 Jan 04; 27(6):685-692.

[99] Hao H. The development of online doctor reviews in China: An analysis of the largest online doctor review website in China. J Med Internet Res 2015 Jun 01; 17(6):e134.

[100] Smith R, Lipoff J. Evaluation of dermatology practice online reviews: Lessons from qualitative analysis. JAMA Dermatol 2016 Feb;152(2):153-157.

[101] Daskivich T, Luu M, Noah B, Fuller G, Anger J, Spiegel B. Differences in online consumer ratings of health care providers across medical, surgical, and allied health specialties: Observational study of 212,933 providers. J Med Internet Res 2018 May 09;20(5):e176.

[102] Wallace BC, Paul MJ, Sarkar U, Trikalinos TA, Dredze M. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. J Am Med Inform Assoc 2014 Jun 10;21(6):1098-1103.

[103] Hao H, Zhang K. The voice of Chinese health consumers: A text mining approach to Web-based physician reviews. J Med Internet Res 2016 May 10;18(5):e108.

[104] Hao H, Zhang K, Wang W, Gao G. A tale of two countries: International comparison of online doctor reviews between China and the United States. Int J Med Inform 2017 Mar;99:37-44.

[105] Zhai Z, Liu B, Xu H, Jia P. Clustering product features for opinion mining. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining; 2011 Feb 9; p. 347-354.

[106] Liu Q, Gao Z, Liu B, Zhang Y. Automated rule selection for aspect extraction in opinion mining. In: Proceedings of the 24th International Conference on Artificial Intelligence; 2015 Jul 25; p. 1291-1297.

[107] Polanyi L, Zaenen A. Contextual valence shifters. In: Computing Attitude and Affect in Text: Theory and Applications. The Information Retrieval Series, vol. 20. Dordrecht: Springer; 2006:1-10.

[108] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory; 1992 Jul 27; p. 144-152.

[109] Tesnière L. Éléments de Syntaxe Structurale. Paris: Klincksieck; 1959.

[110] Matsumoto S, Takamura H, Okumura M. Sentiment classification using word sub-sequences and dependency sub-trees. In: Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining; 2005 May 18; p. 301-311.

[111] Rivas R, Montazeri N, Le NXT, Hristidis V. Github. DTC classifier. URL: https://github.com/rriva002/DTC-Classifier.

[112] Agarwal B, Poria S, Mittal N, Gelbukh A, Hussain A. Concept-level sentiment analysis with dependency-based semantic parsing: A novel approach. Cogn Comput 2015 Jan 20;7(4):487-499.

[113] Wawer A. Towards domain-independent opinion target extraction. Institute of Electrical and Electronics Engineers. In: Proceedings of the 2015 IEEE International Conference on Data Mining; 2015 Nov 14; p. 1326-1331.

[114] Pak A, Paroubek P. Text representation using dependency tree subgraphs for sentiment analysis. In: Proceedings of the 16th International Conference on Database Systems for Advanced Applications (DASFAA 2011); 2011 April 22; p. 323-332.

[115] Kennedy G, McCollough A, Dixon E, Bastidas A, Ryan J, Loo C, et al. Hack harassment: Technology solutions to combat online harassment. In: Proceedings of the First Workshop on Abusive Language Online; 2017 Jul 30; p. 73-77.

[116] Gambäck B, Sikdar UK. Using convolutional neural networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online; 2017 Jul 30; p. 85-90.

[117] Mikolov T, Chen K, Corrado G, Dean J. Arxiv. 2013 Sep 07. Efficient estimation of word representations in vector space. URL: https://arxiv.org/abs/1301.3781.

[118] Lix L, Munakala SN, Singer A. Automated classification of alcohol use by text mining of electronic medical records. Online J Public Health Inform 2017 May 2; 9(1):e069.

[119] Yimam SM, Gurevych I, de Castilho RE, Biemann C. WebAnno: A flexible, Web-based and visually supported system for distributed annotations. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics; 2013 Aug 4; p. 1-6.

[120] Chambers N, Cer D, Grenager T, Hall D, Kiddon C, MacCartney B, et al. Learning alignments and leveraging natural logic. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing; 2007 Jun 28; p. 165-170.

[121] Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. The Stanford CoreNLP natural language processing toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; 2014 Jun 22; p. 55-60.

[122] Chen D, Manning CD. A fast and accurate dependency parser using neural networks. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing; 2014 Oct 25; p. 740-750.

[123] Chang C, Lin C. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2011 Apr 1;2(3):1-27.

[124] Britz D. WildML. 2015 Dec 11. Implementing a CNN for text classification in TensorFlow. URL: http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/.

[125] Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks; 2010 May 19; p. 45-50.

[126] Le Q, Mikolov T. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning; 2014 Jun 14; p. 1188-1196.

[127] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics; 2004 Jul 21; p. 271-278.

[128] Klein D, Manning CD. Fast exact inference with a factored model for natural language parsing. In: Proceedings of the 15th International Conference on Neural Information Processing Systems; 2002 Dec 9; p. 3-10.

[129] Joty S, Carenini G, Ng R, Mehdad Y. Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics; 2013 Aug 4; p. 486-496.

[130] Rivas R, Sadah SA, Guo Y, Hristidis V. Classification of health-related social media posts: Evaluation of post content classifier models and analysis of user demographics. JMIR Public Health and Surveillance. 2020;6(2):e14952.

[131] de Lira VM, Macdonald C, Ounis I, Perego R, Renso C, Times VC. Event attendance classification in social media. Information Processing & Management. 2019 May 1;56(3):687-703.

[132] Ahmad S, Asghar MZ, Alotaibi FM, Awan I. Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. Human-centric Computing and Information Sciences. 2019 Dec 1;9(1):24.

[133] Shen S, Murzintcev N, Song C, Cheng C. Information retrieval of a disaster event from cross-platform social media. Information Discovery and Delivery. 2017 Nov 20;45(4):220-226.

[134] Balsamo D, Bajardi P, Panisson A. Firsthand opiates abuse on social media: Monitoring geospatial patterns of interest through a digital cohort. In: Proceedings of the 2019 World Wide Web Conference; 2019 May 13; p. 2572-2579.

[135] Rao J, Yang W, Zhang Y, Ture F, Lin J. Multi-perspective relevance matching with hierarchical convnets for social media search. In: Proceedings of the 2019 AAAI Conference on Artificial Intelligence; 2019 Jul 17; p. 232-240.

[136] Ruiz EJ, Hristidis V, Ipeirotis PG. Efficient filtering on hidden document streams. In: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media; 2014 May 16.

[137] Li R, Wang S, Chang KC. Towards social data platform: Automatic topic-focused monitor for Twitter stream. Proceedings of the VLDB Endowment. 2013 Sep 1; 6(14):1966-1977.

[138] Li C, Xing J, Sun A, Ma Z. Effective document labeling with very few seed words: A topic model approach. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management; 2016 Oct 24; p. 85-94.

[139] Wang S, Chen Z, Liu B, Emery S. Identifying search keywords for finding relevant social media posts. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence; 2016 Mar 5; p. 3052-3058.

[140] Sadri M, Mehrotra S, Yu Y. Online adaptive topic focused tweet acquisition. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management; 2016 Oct 24; p. 2353-2358.

[141] Proskurnia J, Mavlyutov R, Castillo C, Aberer K, Cudré-Mauroux P. Efficient document filtering using vector space topic expansion and pattern-mining: The case of event detection in microposts. In: Proceedings of the 2017 ACM Conference on Information and Knowledge Management; 2017 Nov 6; p. 457-466.

[142] Li C, Zhou W, Ji F, Duan Y, Chen H. A deep relevance model for zero-shot document filtering. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018 Jul; p. 2300-2310.

[143] Croft WB, Metzler D, Strohman T. Search Engines: Information Retrieval in Practice. Boston, MA, USA: Addison-Wesley; 2010 Feb.

[144] Goudjil M, Koudil M, Bedda M, Ghoggali N. A novel active learning method using SVM for text classification. International Journal of Automation and Computing. 2018 Jun 1;15(3):290-298.

[145] Zhang Y, Lease M, Wallace BC. Active discriminative text representation learning. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence; 2017 Feb 12; p. 3386-3392.

[146] Smailović J, Grčar M, Lavrač N, Žnidaršič M. Stream-based active learning for sentiment analysis in the financial domain. Information Sciences. 2014 Nov 20; 285:181-203.

[147] Pohl D, Bouchachia A, Hellwagner H. Batch-based active learning: Application to social media data for crisis management. Expert Systems with Applications. 2018 Mar 1;93:232-244.

[148] Zhang Y, Zhao P, Cao J, Ma W, Huang J, Wu Q, Tan M. Online adaptive asymmetric active learning for budgeted imbalanced data. In: Proceedings of the 24th ACM International Conference on Knowledge Discovery & Data Mining; 2018 Jul 19; p. 2768-2777.

[149] Li X, Liu B. Learning to classify texts using positive and unlabeled data. In: Proceedings of IJCAI 2003; 2003 Aug 9; p. 587-592.

[150] Li H, Liu B, Mukherjee A, Shao J. Spotting fake reviews using positive-unlabeled learning. Computación y Sistemas. 2014 Sep;18(3):467-475.

[151] Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining; 2008 Aug 24; p. 213-220.

[152] Thorndike RL. Who belongs in the family? Psychometrika. 1953 Dec 1;18(4): 267-276.

[153] Bissoyi S, Mishra BK, Patra MR. Recommender systems in a patient centric social network—a survey. In: Proceedings of the 2016 International Conference on Signal Processing, Communication, Power and Embedded System; 2016 Oct 3; p. 386-389.

[154] Misra R. ResearchGate. News category dataset. June 2018. URL: https://www.researchgate.net/publication/332141218_News_Category_Dataset.

[155] Kullback S, Leibler RA. On information and sufficiency. The Annals of Mathematical Statistics. 1951 Mar 1;22(1):79-86.

[156] Gruda D, Hasan S. Feeling anxious? Perceiving anxiety in tweets using machine learning. Computers in Human Behavior. 2019 Sep 1;98:245-55.

[157] Giachanou A, Crestani F. Like it or not: A survey of Twitter sentiment analysis methods. ACM Computing Surveys (CSUR). 2016 Jun 30;49(2):1-41.

[158] Wu T, Wen S, Xiang Y, Zhou W. Twitter spam detection: Survey of new approaches and comparative study. Computers & Security. 2018 Jul 1;76: 265-284.

[159] Zheng X, Han J, Sun A. A survey of location prediction on Twitter. IEEE Transactions on Knowledge and Data Engineering. 2018 Sep 1;30(9):1652-1671.

[160] Rawat YS, Kankanhalli MS. ConTagNet: Exploiting user context for image tag recommendation. In: Proceedings of the 24th ACM International Conference on Multimedia; 2016 Oct 1; p. 1102-1106.

[161] Zhang Q, Wang J, Huang H, Huang X, Gong Y. Hashtag recommendation for multimodal microblog using co-attention network. In: IJCAI; 2017 Aug 19; p. 3420-3426.

[162] Ma R, Qiu X, Zhang Q, Hu X, Jiang YG, Huang X. Co-attention memory network for multimodal microblog's hashtag recommendation. IEEE Transactions on Knowledge and Data Engineering. 2019 Aug 1.

[163] Faghri F, Fleet DJ, Kiros JR, Fidler S. VSE++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612. 2017 Jul 18.

[164] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7; p. 815-823.

[165] Wu CY, Manmatha R, Smola AJ, Krahenbuhl P. Sampling matters in deep embedding learning. In: Proceedings of the 2017 IEEE International Conference on Computer Vision; 2017 Oct 22; p. 2840-2848.

[166] Zheng Z, Zheng L, Garrett M, Yang Y, Xu M, Shen YD. Dual-path convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 2020 May 19;16(2): 1-23.

[167] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.

[168] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27; p. 770-778.

[169] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20; p. 248-255.

[170] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555. 2014 Dec 11.

[171] Mithun NC, Panda R, Papalexakis EE, Roy-Chowdhury AK. Webly supervised joint embedding for cross-modal image-text retrieval. In: Proceedings of the 26th ACM International Conference on Multimedia; 2018 Oct 15; p. 1856-1864.

[172] Wang Z, Liu X, Li H, Sheng L, Yan J, Wang X, Shao J. Camp: Cross-modal adaptive message passing for text-image retrieval. In: Proceedings of the 2019 IEEE International Conference on Computer Vision; 2019 Oct 27; p. 5764-5773.

[173] Lee KH, Chen X, Hua G, Hu H, He X. Stacked cross attention for image-text matching. In: Proceedings of the 2018 European Conference on Computer Vision; 2018 Sep 8; p. 201-216.

[174] Mithun NC, Li J, Metze F, Roy-Chowdhury, AK. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: Proceedings of the 2018 ACM International Conference on Multimedia Retrieval; 2018 Jun 11; p. 19-27.

[175] Dong J, Li X, Xu C, Ji S, He Y, Yang G, Wang X. Dual encoding for zero-example video retrieval. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition; 2019 Jun 16; p. 9346-9355.

[176] Wray M, Larlus D, Csurka G, Damen D. Fine-grained action retrieval through multiple parts-of-speech embeddings. In: Proceedings of the 2019 IEEE International Conference on Computer Vision; 2019 Oct 27; p. 450-459.

[177] Liu Y, Albanie S, Nagrani A, Zisserman A. Use what you have: Video retrieval using representations from collaborative experts. arXiv preprint arXiv:1907. 13487. 2019.

[178] Yu Y. Kim J. Kim G. A joint sequence fusion model for video question answering and retrieval. In: Proceedings of the 2018 European Conference on Computer Vision; 2018 Sep 8; p. 471-487.

[179] Zhang B. Hu H, Sha F. Cross-modal and hierarchical modeling of video and text. In: Proceedings of the 2018 European Conference on Computer Vision; 2018 Sep 8; p. 374-390.

[180] Shao D, Xiong Y, Zhao Y, Huang Q, Qiao Y. Lin D. Find and focus: Retrieve and localize video events with natural language queries. In: Proceedings of the 2018 European Conference on Computer Vision; 2018 Sep 8; p. 200-216.

[181] Hendricks LA, Wang O, Shechtman E, Sivic J, Darrell T, Russell B. Localizing moments in video with natural language. In: Proceedings of the 2017 IEEE International Conference on Computer Vision; 2017 Oct 22; p. 5803-5812.

[182] Escorcia V, Soldan M, Sivic J, Ghanem B, Russell B. Temporal localization of moments in video collections with natural language. arXiv preprint arXiv:1907. 12763. 2019.

[183] Paul S, Mithun NC, Roy-Chowdhury AK. Text-based localization of moments in a video corpus. arXiv preprint arXiv:2008.08716. 2020.

[184] Hahn M, Silva A, Rehg JM. Action2vec: A crossmodal embedding approach to action learning. arXiv preprint arXiv:1901.00484. 2019 Jan 2.

[185] Zhu D, Ma Y, Liu Y. DeepAD: A joint embedding approach for anomaly detection on attributed networks. In: Proceedings of the 2020 International Conference on Computational Science; 2020 Jun 3; p. 294-307.

[186] Li C, Cao Y, Hou L, Shi J, Li J, Chua TS. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3; p. 2723-2732.

[187] Xiong B, Bao P, Wu Y. Learning semantic and relationship joint embedding for author name disambiguation. Neural Computing and Applications. 2020 Jun 20: 1-12.

[188] Dhingra B, Zhou Z, Fitzpatrick D, Muehl M, Cohen WW. Tweet2vec: Character-based distributed representations for social media. arXiv preprint arXiv: 1605.03481. 2016 May 11.

[189] Vosoughi S, Vijayaraghavan P, Roy D. Tweet2vec: Learning tweet embeddings using character-level CNN-LSTM encoder-decoder. In: Proceedings of the 39th International ACM Conference on Research and Development in Information Retrieval; 2016 Jul 7; p. 1041-1044.

[190] Müller M, Salathé M, Kummervold PE. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. arXiv preprint arXiv:2005.07503. 2020 May 15.

[191] Rawat YS, Kankanhalli MS. ConTagNet: Exploiting user context for image tag recommendation. In: Proceedings of the 24th ACM International Conference on Multimedia; 2016 Oct 1; p. 1102-1106.

[192] Ma R, Qiu X, Zhang Q, Hu X, Jiang YG, Huang X. Co-attention memory network for multimodal microblog's hashtag recommendation. IEEE Transactions on Knowledge and Data Engineering. 2019 Aug 1.

[193] Matsuo S, Shimoda W, Yanai K. Twitter photo geo-localization using both textual and visual features. In: Proceedings of the 2017 IEEE International Conference on Multimedia Big Data; 2017 Apr 19; p. 22-25.

[194] Abraham L. Fastnode2vec. Zenodo. URL: https://doi.org/10.5281/zenodo. 3902632.

[195] Grover A, Leskovec J. Node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13; p. 855-864.

[196] Sayyadiharikandeh M, Varol O, Yang KC, Flammini A, Menczer F. Detection of novel social bots by ensembles of specialized classifiers. In: Proceedings of the

29th ACM International Conference on Information and Knowledge Management; 2020 Oct 19; p. 2725-2732.

[197] Lau JH, Chi L, Tran KN, Cohn T. End-to-end network for Twitter geolocation prediction and hashing. In: Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); 2017 Nov 27; p. 744-753.

[198] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014 Dec 22.

[199] Giannoulakis S, Tsapatsoulis N. Evaluating the descriptive power of Instagram hashtags. Journal of Innovation in Digital Ecosystems. 2016 Dec 1;3(2):114-29.

[200] Lu J, Yang J, Batra D, Parikh D. Hierarchical question-image co-attention for visual question answering. In: Advances in Neural Information Processing Systems; 2016 Dec 5; p. 289-297.

[201] Yang KC, Varol O, Davis CA, Ferrara E, Flammini A, Menczer F. Arming the public with artificial intelligence to counter social bots. Human Behavior and Emerging Technologies. 2019 Jan;1(1):48-61.

[202] Lee K, Eoff B, Caverlee J. Seven months with the devils: A long-term study of content polluters on Twitter. In: Proceedings of the 2011 International AAAI Conference on Web and Social Media; 2011 Jul 5.

[203] Varol O, Ferrara E, Davis C, Menczer F, Flammini A. Online human-bot interactions: Detection, estimation, and characterization. In: Proceedings of the 2017 International AAAI Conference on Web and Social Media; 2017 May 3.

[204] Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In: Proceedings of the 26th International Conference on World Wide Web Companion; 2017 Apr 3; p. 963-972.

[205] Gilani Z, Farahbakhsh R, Tyson G, Wang L, Crowcroft J. Of bots and humans (on Twitter). In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining; 2017 Jul 31; p. 349-354.

[206] Mazza M, Cresci S, Avvenuti M, Quattrociocchi W, Tesconi M. RTbust: Exploiting temporal patterns for botnet detection on Twitter. In: Proceedings of the 10th ACM Conference on Web Science; 2019 Jun 26; p. 183-192.

[207] Cresci S, Lillo F, Regoli D, Tardelli S, Tesconi M. $ FAKE: Evidence of spam and bot activity in stock microblogs on Twitter. In: Proceedings of the 2018 International AAAI Conference on Web and Social Media; 2018 Jun 15.

[208] Yang KC, Varol O, Hui PM, Menczer F. Scalable and generalizable social bot detection through data selection. In: Proceedings of the 2020 AAAI Conference on Artificial Intelligence; 2020 Apr 3; p. 1096-1103.

[209] Rauchfleisch A, Kaiser J. The false positive problem of automatic bot detection in social science research. PloS One. 2020 Oct 22;15(10):e0241045.